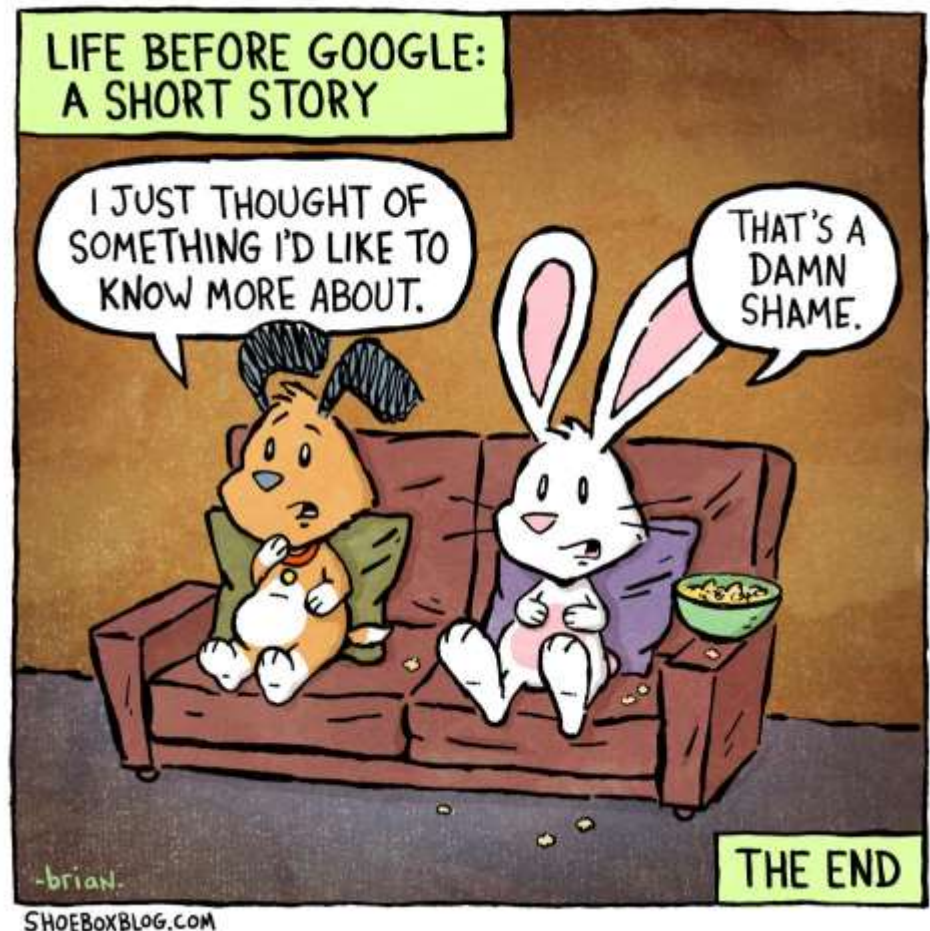


# INF350E / INF384H / CS395T

## Concepts of Information Retrieval (& Web Search)

Matt Lease – [ml@utexas.edu](mailto:ml@utexas.edu) – [ir.ischool.utexas.edu](http://ir.ischool.utexas.edu)





# Matt Lease



Associate Professor  
School of Information  
<http://ir.ischool.utexas.edu>  
[ml@utexas.edu](mailto:ml@utexas.edu)

Talk slides: [slideshare.net/mattlease](https://slideshare.net/mattlease)

## Research Areas

Information Retrieval & Search Engines

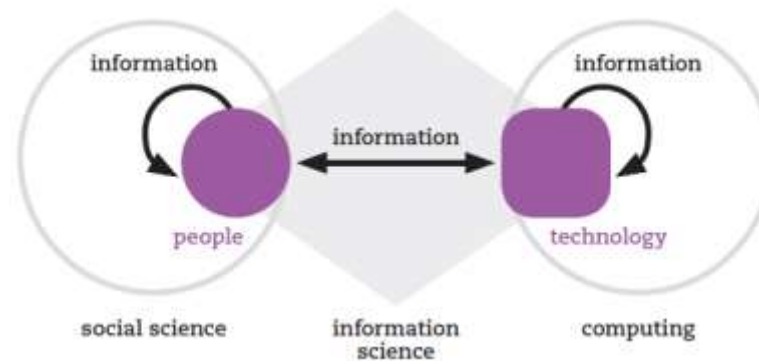
Crowdsourcing & Human Computation

- **NIST TREC 2010-2013:** Ran tracks for the US National Institute of Standards & Technology (NIST) Text REtrieval Conference (TREC)
- **Tutorials:** ACM SIGIR 2011-12 & WSDM 2011, SIAM Data Mining 2013
- **Information Retrieval & Search**
  - Neural Information Retrieval: At the End of the Early Years. *IRJ* 2018
  - Efficient Test Collection Construction via Active Learning. arXiv 2018.
  - ArabicWeb16: A New Crawl for Today's Arabic Web. ACM SIGIR 2016.
- **Human Computation & Crowdsourcing** (e.g., Human-in-the-loop)
  - Why Is That Relevant? Collecting Annotator Rationales for Relevance Judgments. *AAAI HCOMP* 2016.

# What's an Information School?

*“The place where people & technology meet”*

[Wobbrock et al., 2009](#)



*“iSchools”* now exist at over 100 universities around the world



# Human-centered Technology Design



## A GUIDE TO **UX CAREERS**

PRESENTED BY

**ONWARD**  
search

User Experience is one of the fastest growing and most exciting segments in the interactive industry. This guide provides essential information on the different career opportunities within UX, national benchmarks for UX salaries, the hottest metro areas for UX jobs, and tools of the trade for UX professionals.

## **JOB OPPORTUNITIES FOR UX PROFESSIONALS**

**USER  
RESEARCH**

**USABILITY  
ANALYST**

**INFORMATION  
ARCHITECT**

**INTERACTION  
DESIGNER**

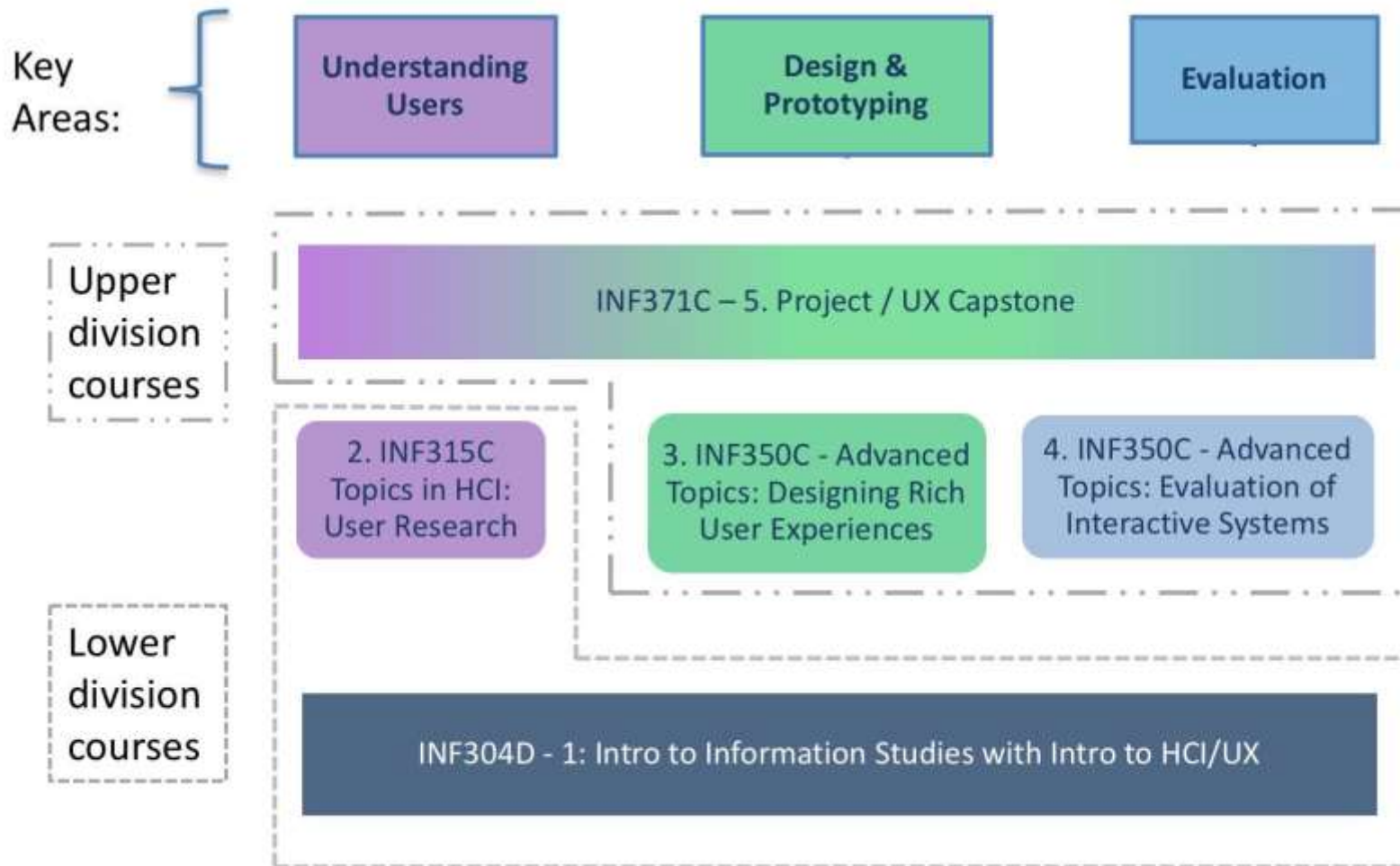
**VISUAL  
DESIGNER**

**UX  
DESIGNER**



# HCI / UX Design - undergraduate minor

HCI - Human Computer Interaction | UX – User Experience



# Information & Computer Science Integrated 5-Year Degree Program

## Bachelors in CS + Masters in IS

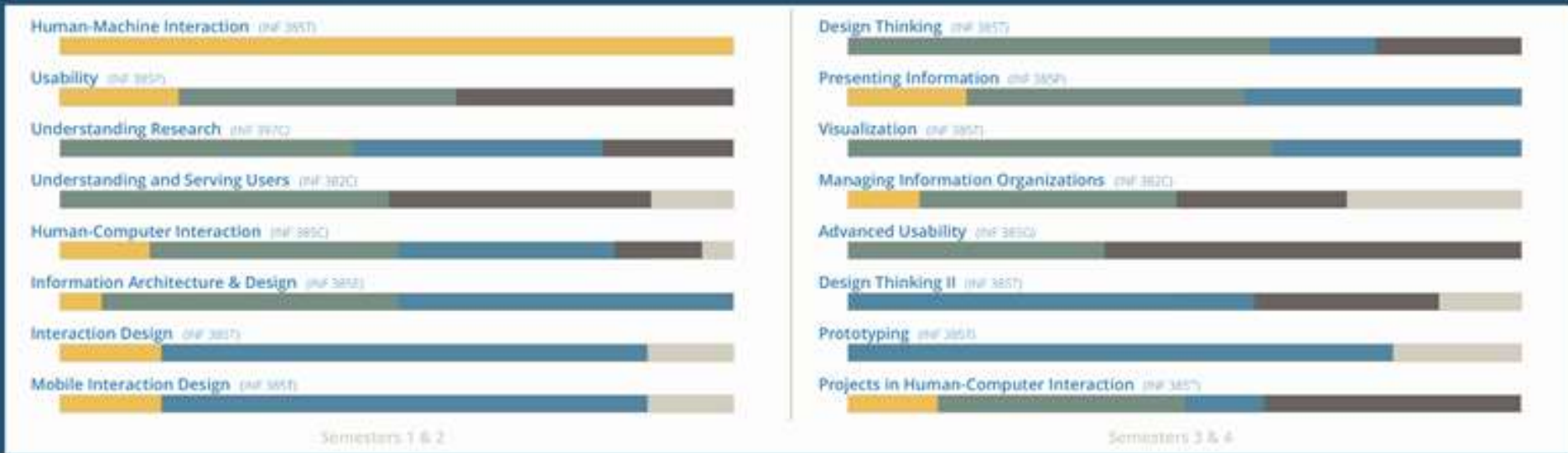




## Empowering our students to be valuable contributors across the UX lifecycle



### UX Track - Course Offerings



# UT Austin “Moonshot” Project

**Goal:** design a future of AI & autonomous technologies that are beneficial — not detrimental — to society.



<http://goodsystems.utexas.edu>



# Fact Checking with Search: Misinformation & Human-AI Partnerships



Matt Lease (University of Texas at Austin)

## “Truthiness” is not a new problem

*“Truthiness is tearing apart our country... It used to be, everyone was entitled to their own opinion, but not their own facts. But that’s not the case anymore.”*

– Stephen Colbert (Jan. 25, 2006)



*“You furnish the pictures and I’ll furnish the war.”*

– William Randolph Hearst (Jan. 25, 1898)

# Information Literacy

*National Information Literacy Awareness Month,  
US Presidential Proclamation, October 1, 2009.*

**“Though we may know how to find the information we need, we must also know how to evaluate it.**

Over the past decade, we have seen a crisis of authenticity emerge. We now live in a world where anyone can publish an opinion or perspective, true or not, and have that opinion amplified...”

Articles tagged: Fake News

304 Total



News > Technology

**Media Firm "Providr" Allegedly Owes Money to Several Angry Facebook Publishers**

17 July 2018 - The company gained prominence by asserting they were compliant with Facebook's rules at a time when few other firms were. They were not, and now their former clients feel duped.



Junk News

**Did Congresswoman Ateesha Nubbins Put an Upper-Age Limit on Voting?**

10 July 2018 - A meme created by a satirical Facebook page convinced some viewers that a fictional congresswoman was attempting to pass a non-existent law.

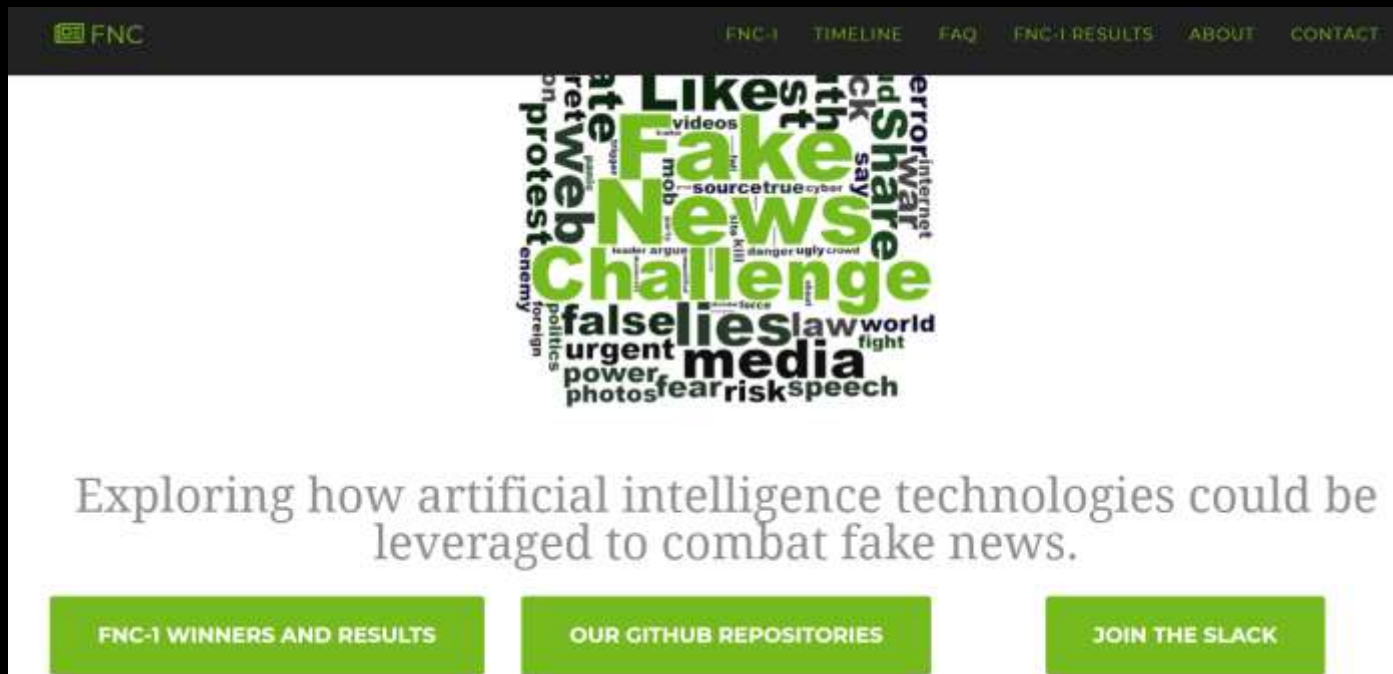


Disney

**Did Disney Announce It Was to Open a Theme Park in Escanaba, Michigan?**



# Automatic Fact Checking



The image shows a screenshot of the FNC (Fact-Checking Network) website. At the top, there is a dark navigation bar with the FNC logo on the left and links for FNC-1, TIMELINE, FAQ, FNC-1 RESULTS, ABOUT, and CONTACT on the right. The main content area features a large word cloud with the most prominent words being "Fake News Challenge", "Lies", "Media", "Share", "Challenge", "Falses", "Law", "World", "Fight", "Urgent", "Power", "Photos", "Fear", "Risk", "Speech", "Like", "Videos", "Source", "True", "Cyber", "Say", "Share", "Internet", "Error", "War", "Mob", "Leader", "Argue", "Danger", "Ugly", "Crowd", "Kill", "Force", "Politics", "Foreign", "Enemy", "Protest", "Web", "Ret", "ate", "Like", "st", "ck", "id", "error". Below the word cloud, the text reads: "Exploring how artificial intelligence technologies could be leveraged to combat fake news." At the bottom, there are three green buttons: "FNC-1 WINNERS AND RESULTS", "OUR GITHUB REPOSITORIES", and "JOIN THE SLACK".

# Design Challenge: How to interact with ML models?

2017 ACM SIGCHI Conference on Human Factors in Computing Systems

## **UX Design Innovation: Challenges for Working with Machine Learning as a Design Material**

**Graham Dove, Kim Halskov**

CAVI, Aarhus University

Aarhus, Denmark

graham.dove@cc.au.dk, halskov@cavi.au.dk

**Jodi Forlizzi, John Zimmerman**

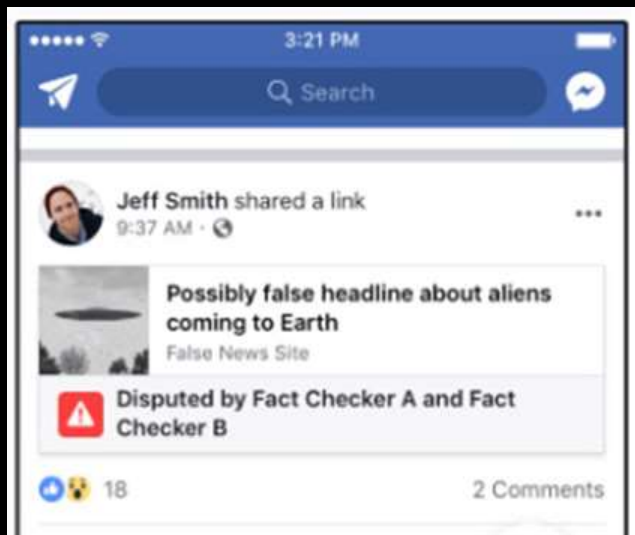
HCII, Carnegie Mellon University

Pittsburgh PA, USA

forlizzi@cs.cmu.edu, johnz@cs.cmu.edu

# Brief Case Study: Facebook

(simpler case: journalist fact-checking)



TECHNOLOGY

## Is There Any Hope for Facebook's Fact-Checking Efforts?

Research is making clear just how hard it is to stop people from believing false stories on social media.

The Atlantic

JON CHRISTIAN SEP 19, 2017



Tessa Lyons, a Facebook News Feed product manager:  
“...putting a strong image, like a red flag, next to an article may actually entrench deeply held beliefs — the opposite effect to what we intended.”



# AI & HCI for Misinformation

“A few classes in ‘use and users of information’ ... could have helped social media platforms avoid the common pitfalls of the backfire effect in their *fake news* efforts and perhaps even avoided ... mob rule, virality-based algorithmic prioritization in the first place.”

<https://www.forbes.com/sites/kalevleetaru/>

Monday, August 5, 2019

The Forbes logo, consisting of the word "Forbes" in a bold, serif font, is displayed within a white rectangular box.

# *Believe it or not: Designing a Human-AI Partnership for Mixed-Initiative Fact-Checking*

Joint work with

An Thanh Nguyen (UT), Byron Wallace (Northeastern), & *more...*



**Matt Lease**

School of Information

University of Texas at Austin



Slides:

[slideshare.net/mattlease](https://slideshare.net/mattlease)



@mattlease

[ml@utexas.edu](mailto:ml@utexas.edu)

# Want to get involved in research?

- Take an Independent Study for credit
- Find a paid Undergraduate Research Assistant job

**EUREKA:** [www.utexas.edu/research/eureka](http://www.utexas.edu/research/eureka)



WHAT STARTS HERE CHANGES THE WORLD  
THE UNIVERSITY OF TEXAS AT AUSTIN



**EUREKA!**

HOME ABOUT PROJECTS

**WHAT RESEARCH TOPIC INTERESTS YOU?**

SEARCH

**Examples:** Bilingualism, Civil War, Education Inequality, Genetics, International Business, Modernism, Urban Policy





# Automatic Fact-Checking

**Given a claim:**

*Facebook Shut Down an AI Experiment Because Chatbots Developed Their Own Language.*

**and relevant article headlines:**

*No, Facebook Did Not Panic and Shut Down an AI Program That Was Getting Dangerously Smart.*

*source: gizmodo.com*

<b>Predict headline stance:</b>	For	Against	Observing
<b>Predict claim veracity:</b>	False	True	Unknown

Predict stance from text features (Ferreira& Vlachos 2016).

Predict veracity from stance+source features (Popat et al. 2017)

# Design Challenges

- Fair, Accountable, & Transparent (AI)
  - Why trust “black box” classifier?
  - How do we reason about potential bias?
  - Do people really only want to know “fact” vs. “fake”?
  - How to integrate human knowledge/experience?
    - Joint AI + Human Reasoning, Correct Errors, Personalization
- How to design strong Human + AI Partnerships?
  - Horvitz, CHI’99: mixed-initiative design
  - Dove et al., CHI’17 *“Machine Learning As a Design Material”*



# Nguyen et al., UIST'18

## Claim Checker

Enter a Claim

[Check Claim](#) [Try A Random Claim](#)

Example Claims:

- Vice Media CEO Shane Smith paid 300,000 for a Las Vegas dinner
- ISIS fighters were caught trying to enter the U.S. via the U.S.-Mexico border
- A N.Y. high schooler earned 72 million in the stock market

## Demo!

---

# Web Search

## Interfaces



# Simple Search Interface Refinements

---

- For “**More results**” requests, stores current ranked list with the user session and displays next set in the list.
- Integrates relevance feedback interaction with “radio buttons” for “NEUTRAL,” “GOOD,” and “BAD” in HTML form.

# Other Search Interface Refinements

---

- Highlight search terms in the displayed document.
  - Provided in cached file on [Google](#).
- Allow for “advanced” search:
  - Phrasal search (“..”)
  - Mandatory terms (+)
  - Negated term (-)
  - Language preference
  - Reverse link
  - Date preference
- Machine translation of pages.

# Web Search Example

- Search suggestions
- Query-biased summarization / snippet generation
- Sponsored search
- Search shortcuts
- Vertical search (news, blog, image)

Web Images Videos Maps News Shopping Gmail more ▾

Google

haiti

- history of haiti
- pictures of haiti
- haitian culture
- haiti climate
- haiti economy
- haiti food
- haiti population
- haiti poverty
- haiti weather
- facts about haiti

- Books
- Finance
- Translate
- Scholar
- Blogs
- YouTube
- Calendar
- Photos
- Documents
- Reader
- Sites
- Groups
- even more »

Search Advanced Search

Results 1 - 10 of about 132,000,000 for haiti. (0.12)

Sponsored Links

- Haiti Earthquake Relief**  
Donate \$25 to Help Children and Families Hurt by the 7.0 Earthquake  
[www.WorldVision.org/Haiti](http://www.WorldVision.org/Haiti)
- Latest News on Haiti**  
Read the latest news about the devastating earthquake. How to help  
[www.SOS-USA.org/HelpHaiti](http://www.SOS-USA.org/HelpHaiti)
- Haiti Earthquake**  
Find Out How To Donate Wisely. Search Tips, Charity Ratings Now!  
[www.CharityNavigator.org](http://www.CharityNavigator.org)
- Earthquake in Haiti**  
IMC sends emergency medical crews. You can help. Donate now.  
[imcworldwide.org](http://imcworldwide.org)
- Global Disasters Maps**  
Access the latest UN information on the world's humanitarian disasters.  
[ReliefWeb.int](http://ReliefWeb.int)
- Aid Haiti Quake Victims**  
Help Habitat respond to the Haiti earthquake. Donate today!  
[www.habitat.org](http://www.habitat.org)
- Haiti News Summary**  
View or sign up to receive news summary, convenient Haiti Reports.  
[www.konpay.org](http://www.konpay.org)
- Haiti Earthquake Appeal**  
Help those most at need.

**CIA - The World Factbook -- Haiti**  
Features map and brief descriptions of the geography, people, government, economy, communications, transportation, military and transnational issues.  
<https://www.cia.gov/library/publications/the-world.../ha.html> - [Cached](#) - [Similar](#)

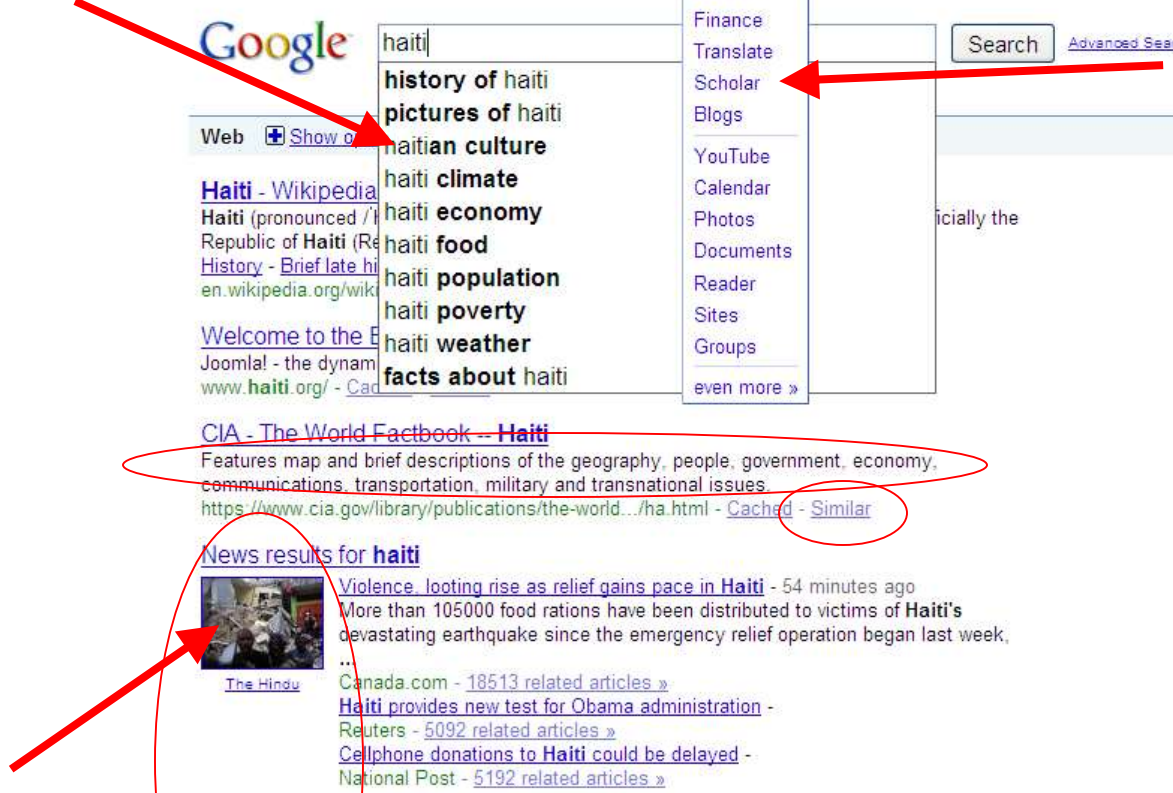
**News results for haiti**

 **Violence, looting rise as relief gains pace in Haiti** - 54 minutes ago  
More than 105000 food rations have been distributed to victims of Haiti's devastating earthquake since the emergency relief operation began last week, ...

[Canada.com - 18513 related articles »](#)  
[Haiti provides new test for Obama administration - Reuters - 5092 related articles »](#)  
[Cellphone donations to Haiti could be delayed - National Post - 5192 related articles »](#)

**Blog posts about haiti**

- [Haiti at Larvatus Prodeo - Larvatus Prodeo - 3 hours ago](#)
- [Commission staff missing in Haiti | Policies | Foreign affairs ... - European Voice - RSS - 6 hours ago](#)
- [Mark Hyman, MD: Haiti Journal: 'Beyond Horror' - The Full Feed from HuffingtonPost.com - 10 hours ago](#)



# Web Search Example

Vertical search (local)

Spelling correction

Personalized search / social ranking



Web [+ Show options...](#) Results

Did you mean: [state of \*\*texas\*\*](#) ←

## [TexasOnline: Official Portal of Texas](#)

Texas Secretary of State Esperanza Hope Andrade is reminding Texans that February 1 is the deadline to register to vote in the 2010 Primary Elections. ...

[www.texasonline.com/](#) - [Cached](#) - [Similar](#) - [Comment](#) [Share](#) [Close](#)

## [TExES](#)

Significant changes to Texas Administrative Code (TAC) §227.10 (a)(3)(C) were approved by the State Board for Educator Certification (SBECE) on October 10, ...

[www.texas.ets.org/](#) - [Cached](#) - [Similar](#) - [Comment](#) [Share](#) [Close](#)

## [In the state of Texas](#)

In the state of Texas, "highly qualified" administrators and teachers are in great demand and in short supply. According to current census information, ...

[texasreview101.com/History.htm](#) - [Cached](#) - [Similar](#) - [Comment](#) [Share](#) [Close](#)

## [Local business results for state of texas near Austin, TX](#) - [Change location](#)



**A** [The Bob Bullock Texas State History Museum: IMAX Theatre](#)  
[www.thestoryoftexas.com](#) - (512) 936-4639 - [50 reviews](#)

**B** [State of Texas: Capitol Visitors Center](#)  
[www.tspb.state.tx.us](#) - (512) 463-0063 - [19 reviews](#)

**C** [State of Texas: General Information](#)

# Web Search Example

The image shows a screenshot of a Yahoo! search results page for the query "haiti". The page layout includes a top navigation bar with tabs for "Web", "Images", "Video", "Local", "Shopping", and "More". The search bar contains the text "haiti" and a yellow "Search" button. Below the search bar, a dropdown menu is open, showing suggestions such as "haiti earthquake", "haiti map", "haiti 90999", "haiti news", and "map of haiti". To the right of the search bar, there are links for "Answers", "Directory", "Jobs", "News", "Sports", and "All Search Services".

On the left side of the page, there is a "Search Pad" section with a "SearchScan - On" indicator and a result count of "946,000,000 results for haiti:". Below this, there are several filters, including "Show All", "Wikipedia", "CNN", "Yahoo! News", "U.S. Department of S...", and "Wikitravel".

The main content area features several sponsored results for "Haiti Earthquake Relief". The first sponsored result is from World Vision, with the headline "Haiti Earthquake Relief" and the text "Donate \$25 to Help Children and Families Hurt by the 7.0 Earthquake." The URL is "www.worldvision.org/haiti". The second sponsored result is from Food for the Poor, with the headline "Haiti Earthquake Disaster" and the text "Donate Now! Help quake victims. Thousands of Haitians dead." The URL is "www.foodforthe poor.org". The third sponsored result is from Habitat for Humanity, with the headline "Aid Haiti Quake Victims" and the text "Habitat for Humanity is working on shelter for victims. Donate today!" The URL is "www.habitat.org/Haiti-Earthquake".

Below the sponsored results, there is a section for "Haiti - Latest News". This section has tabs for "News", "Photos", "Videos", and "Twitter". The first news item is titled "Haiti earthquake: Monday news updates" and is from CNN, dated "10 minutes ago". The text of the article states: "12:20 p.m. Monday, January 18 -- Former Senate Majority Leader Bill Frist, a medical doctor, arrived in Port-au-Prince, Haiti, on Monday to help in the relief... full story". There is a small thumbnail image to the right of this article. Below this, there are two more news items: "U.S. troops boost Haiti aid security as looters swarm" (Reuters via Yahoo! News, 9 minutes ago) and "Why does Haiti suffer so much?" (CNN, 14 minutes ago). The final news item is "More US troops, UN peacekeepers expected for Haiti" (AP via Yahoo! News, 15 minutes ago), with a link to "more Haiti news...".

On the right side of the page, there is another sponsored result for "Haiti Earthquake Relief" with the text "Millions of people affected. Donate now to help us respond." and the URL "www.savethechildren.org/donate". Below this, there is a link that says "See your message here...".



# Web Search Example

Web Images Videos Shopping News Maps More MSN Hotmail

bing

haiti

ALL RESULTS 1-20 of 35,700,000 results - [Advanced](#) Sponsored sites

**HAITI**

- News about Haiti
- Images of Haiti
- Haiti Economy
- Haiti Food
- Haiti Tourism
- Haiti Map
- Haiti Travel
- Reference Articles on Haiti

RELATED SEARCHES

- Haiti Radio
- Haiti Music
- Haiti News Network
- Haiti News Today
- Haiti Sakapfet
- Haiti People
- Police Haiti
- Haiti News Picture

SEARCH HISTORY

Now you can go back further with search history. [Learn More.](#)

[Aid Haiti Quake Victims](#) - [www.Habitat.org/Haiti-Earthquake](http://www.Habitat.org/Haiti-Earthquake) Sponsored sites  
Habitat for Humanity is working on shelter for victims. Donate today!

[Haiti Earthquake Aid](#) - [www.SavetheChildren.org/donate](http://www.SavetheChildren.org/donate)  
Millions of people affected. Donate now to help us respond.

[News about haiti](#)


Rebuilding Haiti  
Haiti's infrastructure for things like clean water and sewage disposal was primitive before last... - World 2 hours ago

Man rescued from Haiti hotel plans to return - WTOP 44 minutes ago  
French minister criticizes US aid role in Haiti - WTOP 2 hours ago

[See today's top stories](#) - [Create news alert](#)

Share this story [f](#) [t](#) [✉](#)

[Images of haiti](#)



[Haiti - Wikipedia, the free encyclopedia](#)  
Haiti (Haitian Creole: Ayiti), officially the Republic of Haiti (République d'Haiti ; Repiblik Ayiti) is a Haitian Creole - and French-speaking Caribbean country. Along with ...  
History - Politics - Departments ... - Geography  
[en.wikipedia.org/wiki/Haiti](http://en.wikipedia.org/wiki/Haiti) - [Wikipedia on Bing](#)

[Help Those in Need Today](#)  
Expect change when you give to The Salvation Army. Make a donation.  
[SalvationArmy.org](http://SalvationArmy.org)  
[See your message here](#)

ABC News Video



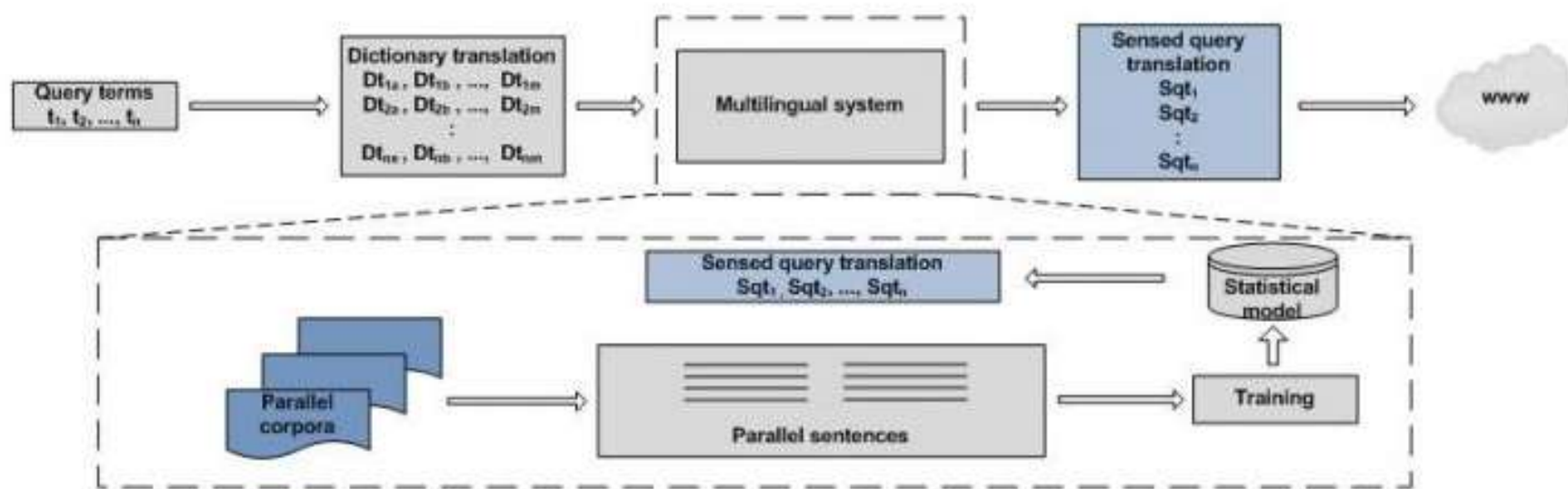
# Cross-Lingual IR

- 2/3 of the Web is in English
- About 50% of Web users do not use English as their primary language
- Many (maybe most) search applications have to deal with multiple languages
  - *monolingual search*: search in one language, but with many possible languages
  - *cross-language search*: search in multiple languages at the same time

# Cross-Lingual IR

## Ideal

- Let user express query in native language
- Search information in multiple languages
- Translate results into user's native language



# Vertical Search

- Aka/related: federated / distributed / specialty
- Searching the “Deep” web
- One-size-fits-all vs. niche search
  - Query formulation, content, usability/presentation



# Clustering Results

---

- Group search results into coherent “clusters”:
  - “microwave dish”
    - One group of on food recipes or cookware.
    - Another group on satellite TV reception.
  - “Austin bats”
    - One group on the local flying mammals.
    - One group on the local hockey team.
- Northern Light used to group results into “folders” based on a pre-established categorization of pages (like DMOZ categories).
- Alternative is to dynamically cluster search results into groups of similar documents.

# Other Visual Interfaces

apple

n.g. 'apple' or 'computer science'

You searched for apple. 116 results and 23 categories

Apple >

- AAPL Stock (24)**  
 Dig - AAPL Stock Tanking  
 AAPL (APPLE INC) Stock analysis and research by investors on UpDown.com  
 AAPL Stock - Wall Street Survivor  
 AAPL Apple Inc. Stock Message Board, News, Social Network  
 Quotes, Activity and Analysis for Apple Inc (NASDAQ:AAPL)  
 Apple Inc. (AAPL) Stock - Seeking Alpha  
 AAPL - Stock Quote for Apple Inc. - MSN Money News...
- Mac News (8)**  
 Macworld  
 Apple Investor Relations Earnings Releases  
 Apple Computer (7)  
 AAPL  
 Apple (AAPL)  
 Apple Leopard Server - Full Review - Reviews by PC Magazine  
 Apple Products (6)  
 Mac Laptops, iPods and Macintosh Computers  
 Apple Online Store  
 Apple - AppleTV  
 www.apple.com/apple

Privatix

Go > http://P-prlage01/staging/office/Collections/2009/Pl...

NFL Players - 2009 Season | Team: Tampa Bay Buccaneers

Sort: Years of Experience

Filter by Keyword

Team

- Buffalo Bills
- New Orleans Saints
- Baltimore Ravens
- Detroit Lions
- Tampa Bay Buccaneers
- Houston Texans
- Kansas City Chiefs
- St. Louis Rams
- Atlanta Falcons
- Cincinnati Bengals
- Green Bay Packers
- Carolina Panthers
- Miami Dolphins

Position

Years of Experience

2008 Salary

Touchdowns

Passing Yards

Rushing Yards

Receiving Yards

Tasks

Browse packages, types, and methods related to keywords

Judge relevancy by example counts

Filter pages with code examples by package, type, or member

output:acrobati - Assieme - Mozilla Firefox

File Edit View History Bookmarks Tools Help

Assieme

output:acrobati Search Start Top

Package	Type	Members
57 examples	com.lowagie.text	- Javadoc
13 examples	org.apache.fop.apps	- Javadoc
6 examples	javax.print.attribute.standard	- Javadoc
31 examples	com.lowagie.text.pdf	- Javadoc
4 examples	com.lowagie.text.rtf	- Javadoc

Filter: com.lowagie.text.pdf

Java meets PDF - Java Magazine - Internet & Enterprise...

http://www.javamagazine.com/online-int...\_speccon.pdf...html

```

java.lang.String com.lowagie.text.pdf.PdfWriter.java.lang.Math.java.lang.Exception
com.lowagie.text.pdf.PdfParser com.lowagie.text.Rectangle.java.awt.Graphics2D
com.lowagie.text.Document.java.io.FileOutputStream.java.awt.Color
[Text jar]
    
```

Text tutorial for Java Studio Creator 2

http://www.kit-indonesia.com/tutorials/text1/

```

java.servlet.http.HttpServlet.java.lang.String.com.lowagie.text.DocumentException
java.servlet.ServletException.java.util.Date.com.lowagie.text.Document
java.servlet.http.HttpServletResponse.com.lowagie.text.pdf.PdfWriter
com.lowagie.text.html.HtmlWriter.java.servlet.http.HttpServletRequest.java.io.IOException
java.io.PipedStream.com.lowagie.text.html.HtmlWriter.com.lowagie.text.Paragraph
[Text jar] [Text jar]
    
```

package contents:

```

import java.awt.*;
import java.io.*;
import com.lowagie.text.*;
import com.lowagie.text.pdf.*;

public class HelloWorld {

    private final static String out = "contented.pdf";
    Document document = new Document(PdfPageSize.A4);

    try {
        PdfWriter.getInstance(document, new FileOutputStream(
            document.open());
        document.add(new Phrase("Hello World"));
    }
    
```

See and download required JAR files

Page summaries show Java types used in code examples

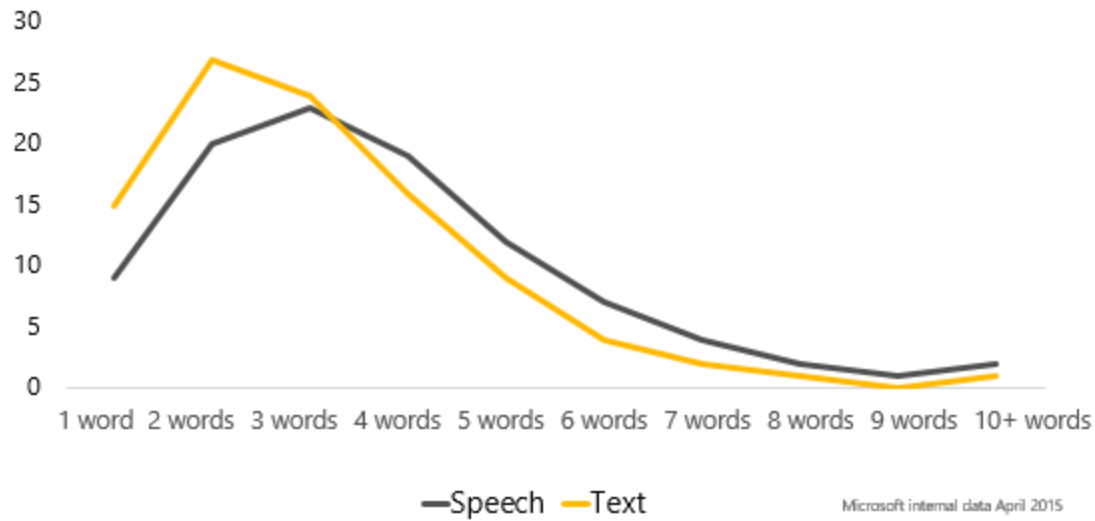
Hover over page titles to see code example previews

Context-sensitive menus at highlighted types (see Figure 2)



# Speech Queries are Longer

---





# User Query Length

- Users tend to enter short queries.
  - Study in 1998 gave average length of 2.35 words.
- Evidence that queries are getting longer.

Percentage of U.S. clicks by number of keywords				
Subject	Jan-08	Dec-08	Jan-09	Year-over-year percent change
1 word	20.96%	20.70%	20.29%	-3%
2 words	24.91%	24.13%	23.65%	-5%
3 words	22.03%	21.94%	21.92%	0%
4 words	14.54%	14.67%	14.89%	2%
5 words	8.20%	8.37%	8.68%	6%
6 words	4.32%	4.47%	4.65%	8%
7 words	2.23%	2.40%	2.49%	12%
8+ words	2.81%	3.31%	3.43%	22%
<i>Note: Data is based on four-week rolling periods (ending Jan. 31, 2009; Dec. 27, 2008; and Jan. 26, 2008) from the Hitwise sample of 10 million U.S. Internet users.</i>				
<b>Source: Hitwise, an Experian company</b>				

# Spoken Search



Longer and more natural queries emerge given support for spoken input [Du and Crestiani'06]

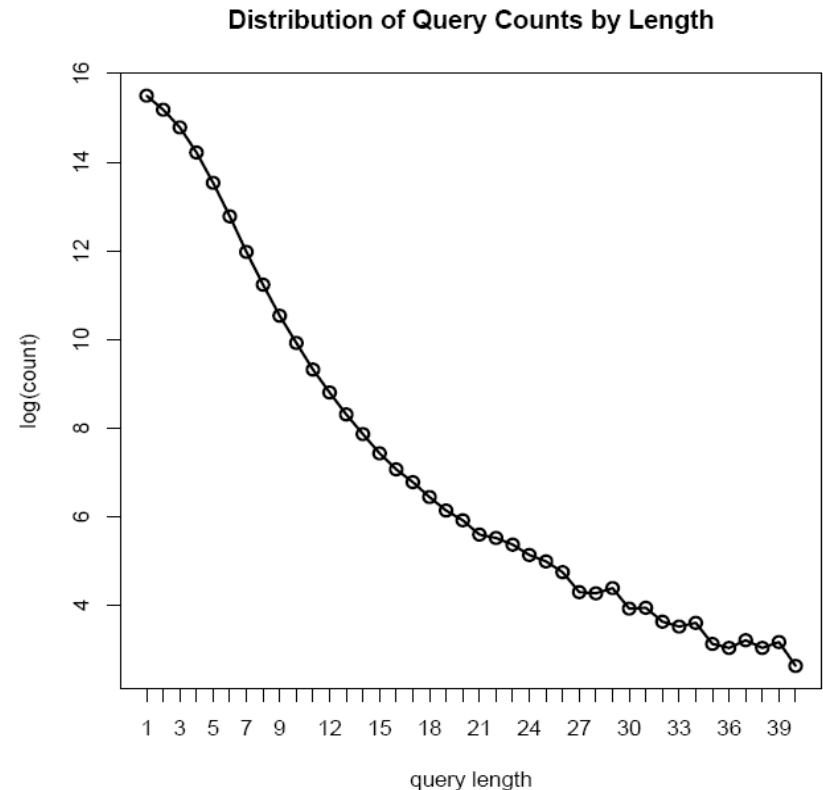
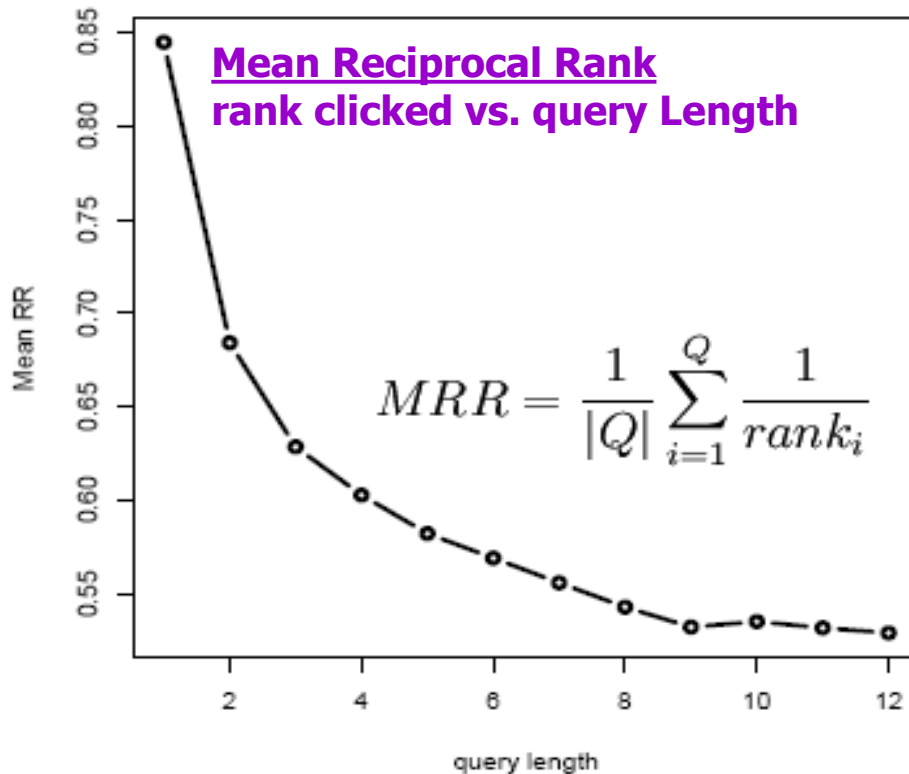
See also: studies by Nick Belkin

# Long / Verbose Web Queries

- User queries from  Live Search



- Analysis by [Bendersky and Croft'09]



# Spoken "Document" Retrieval

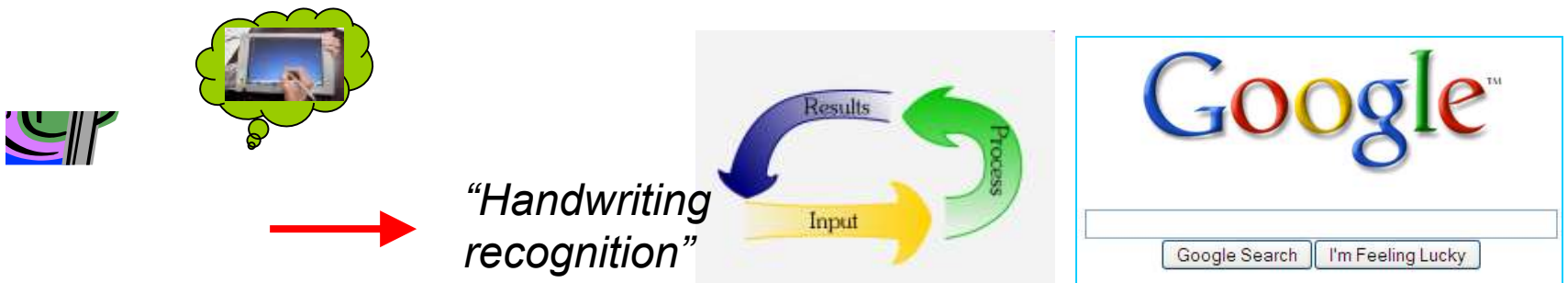


# User <-> Search Engine Feedback Cycle

Query formulation reflects an ongoing dialog between users and search engines

- Users formulate queries *for the* search engine, based on a mental model of what it “understands”
- Search engines optimize their “understanding” *for the* (most frequent) submitted queries
- Individual session and long term, personal and aggregate

Result: query “language” is continually evolving



# Verbosity and Complexity

- Complex information requires complex description
  - Information theory [Shannon'51]
  - Human discourse implicitly respects this [Grice'67]
- Simple searches easily expressed in keywords
  - navigation: “alaska airlines”
  - information: “american revolution”
- Verbosity naturally increases with complexity
  - More specific information needs [Phan et al.'07]
  - Iterative reformulation [Lau and Horvitz'99]



# Blog Search


**Technorati** beta Blogs Posts Search for posts... Join

Technology Business Entertainment Lifestyle Sports Politics Videos Blogging **Twitterati**

Blog Directory Top 100 Tags People Technorati Blog Write for Technorati State of the Blog

Ads by Google Blog Directory Video Blog Blogger Photos Blog Gratuito


### Today on Technorati



**EU Joins US in Pledge To Rebuild Haiti**

The European Union has joined the United States in its pledge to rebuild the broken and devastated country of Haiti.


[Read more](#) in Politics



**Amid Death, Chaos and Fear Heroes Emerge**

With odds against him, Haitian doctor Claude Surena turns his home into a makeshift hospital to treat the injured.


[Read more](#) in Lifestyle



**The Severe Consequences of Defaulting On Your Mortgage**

With so many people defaulting on their mortgages, the negative financial impact is too often ignored.

[Read more](#) in Business



**American Idol Enters the Next Round of Auditions**

"Pants On The Ground" is already an internet sensation. What will the next round of auditions have to offer? Show airs Tuesday.

[Read more](#) in Entertainment

### Top Blogs

[View the entire blog directory](#)

Top 5 risers		Top 5 fallers	
53	<b>SignOnSanDiego.com:</b> Authority: 804	20	
87	<b>The Plum Line</b> Authority: 786	13	
85	<b>SlashGear</b> Authority: 787	-18	
79	<b>Ezra Klein</b> Authority: 793	-12	

**Google blogs** haiti Search Blogs

### Blog results

[Browse Top Stories](#)

Published

- [Last hour](#)
- [Last 12 hours](#)
- [Last day](#)
- [Past week](#)
- [Past month](#)
- [Anytime](#)
- [Choose Dates](#)

Subscribe:

- [Blogs Alerts](#)
- [Atom](#) | [RSS](#)

Related Blogs: [The Livesay \[Haiti\] Weblog](#) - <http://livesayhaiti.blogspot.com/>  
[Haiti Innovation | Choice, Partnership, Community](#) - <http://www.haitiinnovation.org/>  
[Northwest Haiti Christian Mission](#) - <http://www.nwhcm.org/>  
[Haiti Children | Mercy & Sharing | Donate to help Haiti](#) - <http://www.haitichildren.com/>  
[Canada Haiti Action](#) - <http://canadahaitiaction.ca/>

[Latest Updates on the Crisis in Haiti - The Lede Blog - NYTimes.com](#)  
 4 hours ago by By ROBERT MACKEY  
 On Monday The Lede is continuing to supplement reporting by our colleagues aftermath of Tuesday's earthquake by pointing to news and information that are encouraged to share any first-hand accounts they ...  
[The Lede](#) - <http://thelede.blogs.nytimes.com/> - [References](#)  
[\[ More results from The Lede \]](#)

[Haiti at Larvatus Prodeo](#)  
 4 hours ago by Mark  
 Blogging politics, culture, sociology and life from Brisvegas.  
[Larvatus Prodeo](#) - <http://larvatusprodeo.net/>

[Commission staff missing in Haiti | Policies | Foreign affairs](#)  
 7 hours ago  
 commission staff missing in haiti three members of the eu's delegation missing after last weekend's devastating earthquake.  
[European Voice](#) - [RSS](http://www.europeanvoice.com/) - [References](#)  
[\[ More results from European Voice - RSS \]](#)

[Art to Support Haiti | Abduzeedo | Graphic Design Inspiration](#)  
 5 hours ago by paul0v2  
 After the catastrophic earthquake that struck Haiti last Tuesday, everyone however they can and the design community is also trying to do it's part James White, Chuck Anderson, Nathalie Bertin, ...  
[Abduzeedo | Graphic Design Inspiration](#) - <http://abduzeedo.com/>



# μ-Blog Search (e.g. Twitter)



## Twittorati

Where the blogosphere and twittersphere meet  
A Technorati™ site powered by Muck Rack



Search Twittorati



Tweets

Latest Twittorati Chatter

Top Links

Most Popular Links

Top Blogs

The Twittorati Top 100

Latest Photos

See What Everyone's Chatting About

Hide Panels - Clear Page - Pause Tweets - Link here  
Empty Queue    Queued Tweets: 52    New Tweet

**Trends**

- All Terms
- #thingsyouneverst
- Shorty Award
- #mm
- #Donttalktome
- Happy MLK Day
- #musicmonday
- #nowplaying
- Martin Luther King
- MLK
- #IHaveaDream

**Lists**

All Lists

haiti  Add

That wasn't a valid list name, the correct form is @usernamefirstname. For example: @twitteam [Remove Alert](#)

**Searches**

All Searches

haiti  Add

**Geolocation**

Community RT @theylilous; RT @marcoponce: U.S. takes control of Haiti <http://bit.ly/66c7vH> More evidence of the quake being man made 4 man made p ...

**S** SavannahNow VIDEO: Raw Video: Former President Clinton Arrives in Haiti. Former President Bill Clinton and daughter Chelsea ar... <http://bit.ly/6dtWQZ>

WardellJClark: "If Martin Luther was living He Wouldnt Let This Be!" Michael Jackson very appropo. for today and Haiti and the World in General

betriesignolcom RT @mbait: Justin Timberlake, Bono & Alicia Keys For Haiti Event <http://bit.ly/6cADr>

secaningscity RT @adzansari: Woops. Got wasted last night and drunk texted \$15,000 in donations to Haiti relief efforts.

allbran Just donated to the Haiti cause. Canada government will match all donations up to \$50. Very nice of them too.

LilacGirl2 RT @CruiseCriticUK; RT @CruiseEditor: Support for Royal Caribbean's call Friday at Haiti pretty strong on both @CruiseCriticUK & @cruise ...

demajeannerter RT @jacktegerstein: Aggregate of #Haiti Earthquake Lesson Plans <http://bit.ly/6y4g5e> Please let me know of others

RIA\_Novest5 #news : Haiti gives a chance to America and the rest of the world <http://bit.ly/87sADR>

norabf #Haiti RT @shartheoudous: Ppl in Leogane need food desperately.UN won't come until 'security is confirmed' <http://twitpic.com/yog2d>

# Book Search

- Find books or more focused results
- Detect / generate / link table of contents
- Classification: detect genre (e.g. for browsing)
- Detect related books, revised editions
- Challenges
  - Variable scan quality, OCR accuracy
  - Copyright
  - Monetary model

# Other IR-Related Tasks

---

- Automated document categorization
- Information filtering (spam filtering)
- Information routing
- Automated document clustering
- Recommending information or products
- Information extraction
- Information integration
- Question answering

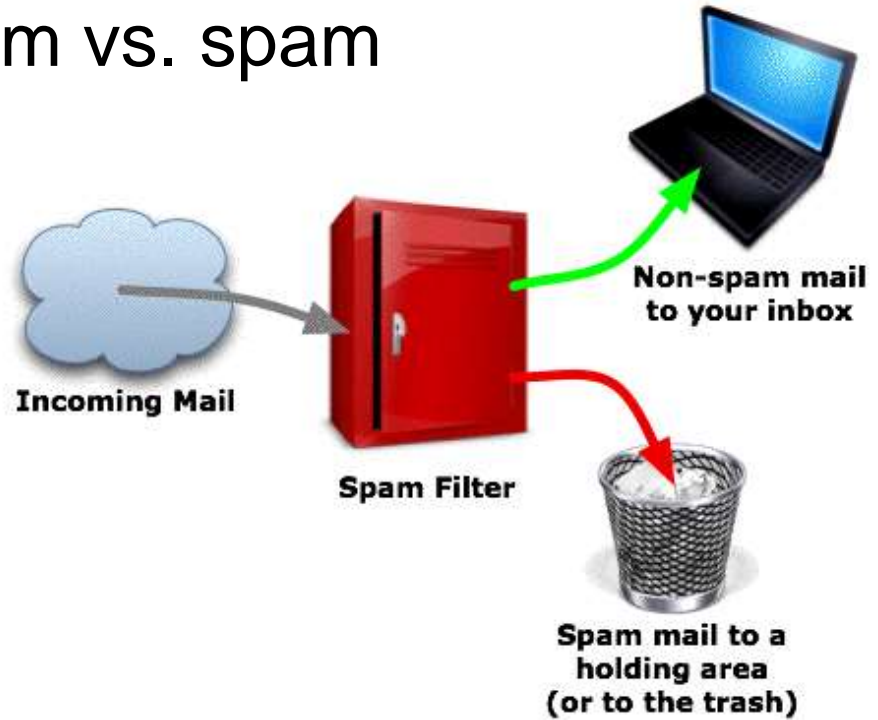
# Dimensions of IR

<b>Content</b>	<b>Applications</b>	<b>Tasks</b>
Text	Web search	Ad hoc search
Images	Vertical search	Filtering
Video	Enterprise search	Classification
Scanned docs	Desktop search	Question answering
Audio	Forum search	
Music	P2P search	
	Literature search	

---

# Routing / Filtering

- Given standing query, analyze new information as it arrives
  - Input: all email, RSS feed or listserv, ...
  - Typically classification rather than ranking
  - Simple example: Ham vs. spam
  - Anomaly detection



# Collaborative Search



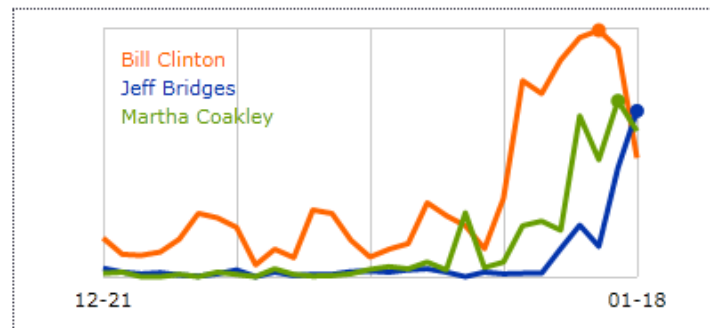
# Entity Search



[What's EntityCube?](#)

**People:** [Barack Obama](#) [James Cameron](#) [Scott Brown](#) [Jay Leno](#) [Martha Coakley](#) ▶ [More Popular Names](#)

**Keywords:** [Best Director](#) [Beijing](#) [Microsoft](#) [Joe the plumber](#) [Australian Open](#) [Economic Downturn](#)



[Sports](#) ▶ [More..](#)

**Scott Brown** Powered by Social Strata 32 mins ago

**Brett Favre** More troops, aid go to Haiti, but hunge... 32 mins ago

**Maria Sharapova** The Daily Mail: AUSTRALIAN OPEN 2... 8 mins ago

**Justine Henin** Clijsters brightens up dreary day at... 39 mins ago

**Tony Romo** Three Dominant Home Teams " and the Jets 30 mins ago

[Entertainment](#) ▶ [More..](#)

**James Cameron** Cameron's 'Avatar' wins top honours ... 27 mins ago

**Jay Leno** SNL's 'Weekend Update' explains Leno-O'Brien... 34 mins ago

**Conan O'Brien** Real Madrid stud Cristiano Ronaldo i... 34 mins ago

**George Clooney** Taylor Lautner: Golden Globes 2010 ... 38 mins ago

**John Denver** Man in custody after standoff in West B... 39 mins ago



# Expertise Search

EXPERTSEARCH.CO.UK

There are **9** records matching your search request :

**Area of Expertise = Information Retrieval**

Your search took **0.250 seconds** to perform.

Name: [Mr John Allcock](#)

[web site](#)



Town/County: London

Organisation: Bristows, Solicitors

Occupation: Solicitor and European  
Patent/Trade Mark Attorney

Name: [Mr Matthew J Atha](#)

[web site](#)



EWI

THE ACADEMY EXPERTS

Town/County: Wigan, Lancs

Organisation: Independent Drug Monitoring  
Unit (IDMU)

Occupation: Drug Abuse Research &  
Information Consultant

Name: [Miss Annette Clarey](#)

[web site](#)



Town/County: Slough, Berks

Organisation: BioMark Forensics Ltd

Occupation: Forensic Biologist



BIO-MARK FORENSICS

Name: [Mr Andrew Fox](#)

[web site](#)



Town/County: Plymouth, Devon

Organisation: Audax Digital Forensic

Occupation: Computer Forensic  
Consultant



Prospective Students Business and Government Current Students Alumni Faculty

[College of Engineering](#) >> [Academic Programs](#) >> [Graduate School](#) >>

## Faculty Expertise -- Search

This faculty search tool can be used to identify faculty members within the College of Engineering who have expertise in specific areas of interest. This can be useful for identifying

## Auburn Engineering Faculty Search Engine

Computer Science and Software Engineering

information retrieval

Search

JUAN E. GILBERT

TSYS Distinguished Associate Professor  
Computer Science and Software Engineering  
3101 Shelby Center  
Phone: (334) 844-6316 Fax: (334) 844-6329  
E-mail: [gilbert@auburn.edu](mailto:gilbert@auburn.edu)  
Website: <http://www.juangilbert.com>

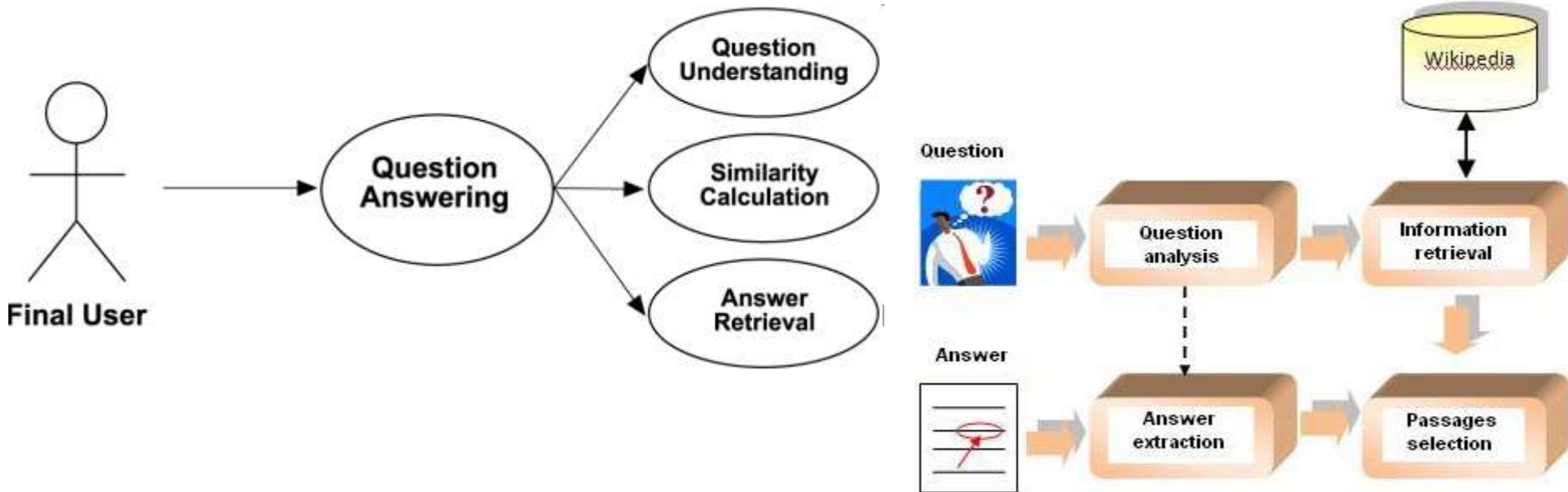
- Human-Centered Computing, Human-Computer Interaction, Spoken Language Systems, Databases, Information Management, Advanced Learning Technologies, Ethnocomputing

W. H. CARLISLE

Associate Professor  
Computer Science and Software Engineering  
110 Dunstan Hall  
Phone: (334) 844-6308 Fax: (334) 844-6329  
E-mail: [carlisle@auburn.edu](mailto:carlisle@auburn.edu)

- Languages and algorithms for cooperative autonomous systems, distributed processing, and distributed information sharing and system management.

# Question Answering & Focused Retrieval



## AnswerBus

what is information retrieval

Ask

Type in your question in English, French, Spanish, German, Italian or Portuguese.

n

Question:

what is information retrieval

Possible answers: [XML](#) [TXT](#)

1. There is a common confusion, however, between data retrieval, document retrieval, information retrieval, and text retrieval, and each of these its own bodies of literature, theory, praxis and technologies.
2. Information Retrieval (IR) is a discipline of studying theories, models, and techniques that deal with the representation, storage, organization, an retrieval of information items so that they can be useful to humans.

# Community QA

**YAHOO!** ANSWERS

## News & Events

### Did bill clinton really engineer this economic

1 ☆ In [Current Events](#) - Asked by [eja123](#) - 7 answers - 45 minutes ago

### Why did Bill o'reilly quit his daily radio show?

☆ In [Media & Journalism](#) - Asked by [davidagoldsmith](#) - 1 answer - 1 minute ago

### Why do people think Obama won cuz he's black?

☆ In [Current Events](#) - Asked by [jade3712](#) - 7 answers

### How many people did Saddam Hussein kill?

☆ In [Current Events](#) - Asked by [eja123](#) - 2 answers

### Is obama going to improve foreign relations?

☆ In [Current Events](#) - Asked by [eja123](#) - 4 answers

### Do You Think Chris McCandless "Deserved" to Die?

☆ In [Media & Journalism](#) - Asked by [veqqieiss!](#) - 0 answers - 0 minutes ago

### Obama haters, were you satisfied with the job the did the past 8 years?

☆ In [Current Events](#) - Asked by [Mallory](#) - 5 answers

### What qualities should a Journalist have?

1 ☆ In [Media & Journalism](#) - Asked by [eja123](#) - 0 answers

 WikiAnswers.com 10,000,000 Qs

## Politics and Society

[First page] 2 3 4 5 6 7 8 9 10 11 12 Next > [Last Politics and Society page]

### Why is Latin America called 'Latin' America? [Edit categories]

Popularity: 335

### What is the difference between Saudi Arabia and Arabia? [Edit categories]

Popularity: 76

### What were the Maya achievements? [Edit categories]

Popularity: 48

 wondir



Science

11 Mar '09, 00:56 (SCI) What is the scientific name for dragonfly ants? ( 1 response )

10 Mar '09, 14:40 (SCI) what are three uses of boron chemicals ( 1 response )

10 Mar '09, 09:12 (SCI) Does nucleus only contain protons and neutrons? ( 2 responses )

10 Mar '09, 02:16 (SCI) what colour is violet????????? ( 3 responses )

10 Mar '09, 02:00 (SCI) how has globalisation had an impact on the role of the state and the concept of sovereignty

10 Mar '09, 01:48 (SCI) is s. equi sub specie equi an aerobic bacteria?

10 Mar '09, 00:01 (SCI) when did the sugar act occur? ( 1 response )

09 Mar '09, 16:16 (SCI) Does it rain sperm at a baby shower? ( 3 responses )

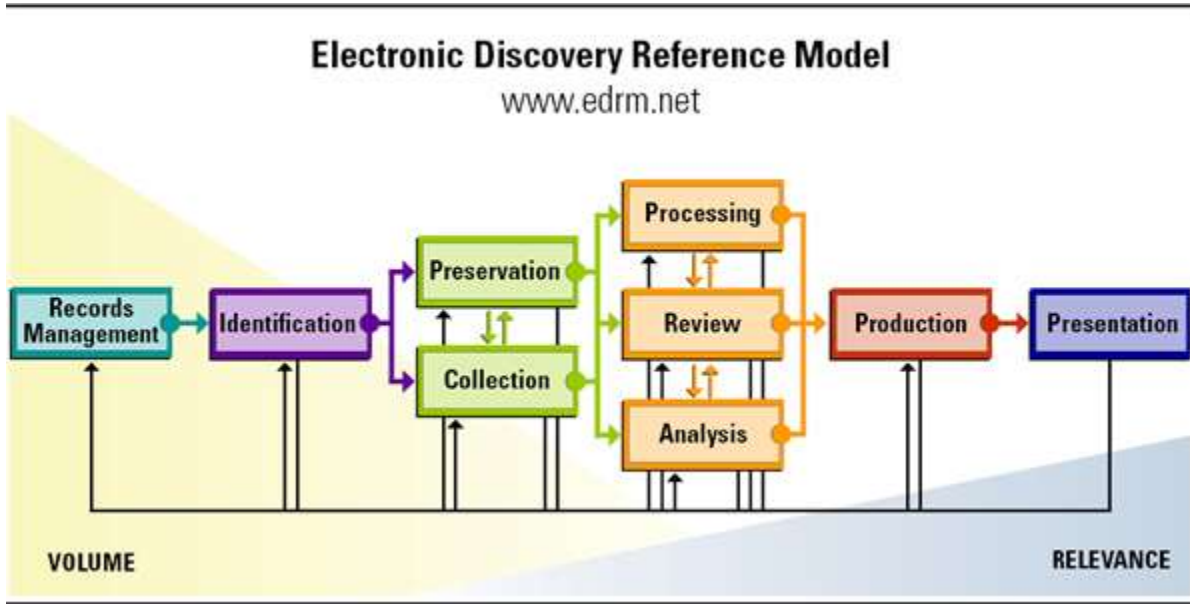
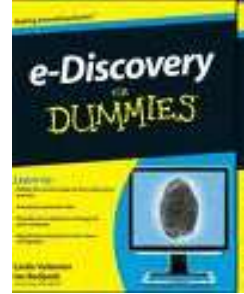
09 Mar '09, 09:30 (SCI) How much stronger is the new vitreous carbon material invented by the Tokyo Institute of Technology compared with the material made from cellulose?

08 Mar '09, 22:36 (SCI) where can i go to find answers about alcohol related questions? ( 2 responses )

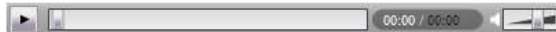
07 Mar '09, 22:43 (SCI) if the earth suddenly stopped spinning ,would we go flying through the air? ( 4 responses )



# e-Discovery



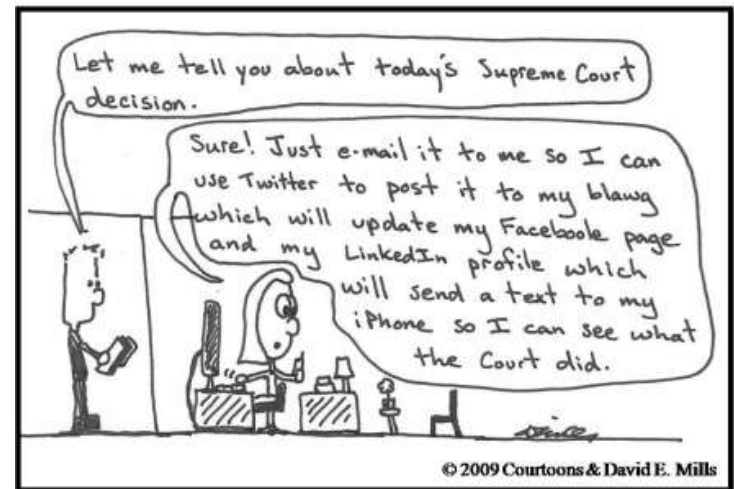
## Electronic Discovery Lessons from Dedicated Review Teams



Podcast: [Play in new window](#) | [Download](#)

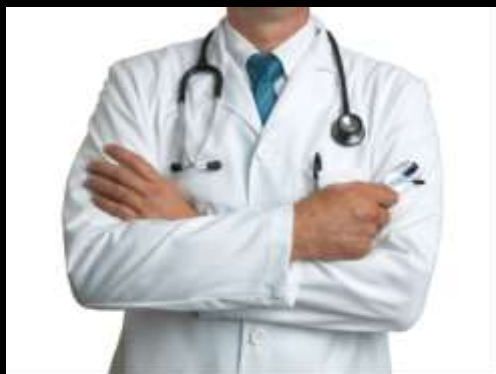


Listen to Matt Clarke, a shareholder and member of the Ryley Carlock & Applewhite's Document Control Group, and Karl Schieneman, Director of Legal Analytics & Review at Jurinnov as moderator, as they discuss working with dedicated review spaces for staffing electronic discovery projects. Matt is the



How lawyers use technology to make things casier.

# Systematic Review is e-Discovery in Doctor's Clothing



Joint work with

Gordon V. Cormack (U. Waterloo)	An Thanh Nguyen (U. Texas)
Thomas A. Trikalinos (Brown U.)	Byron C. Wallace (U. Texas)

SIGIR 2016 Workshop on Medical IR (MedIR)



New

OLD

RCT

Franklin

7-12

7-12

7-12



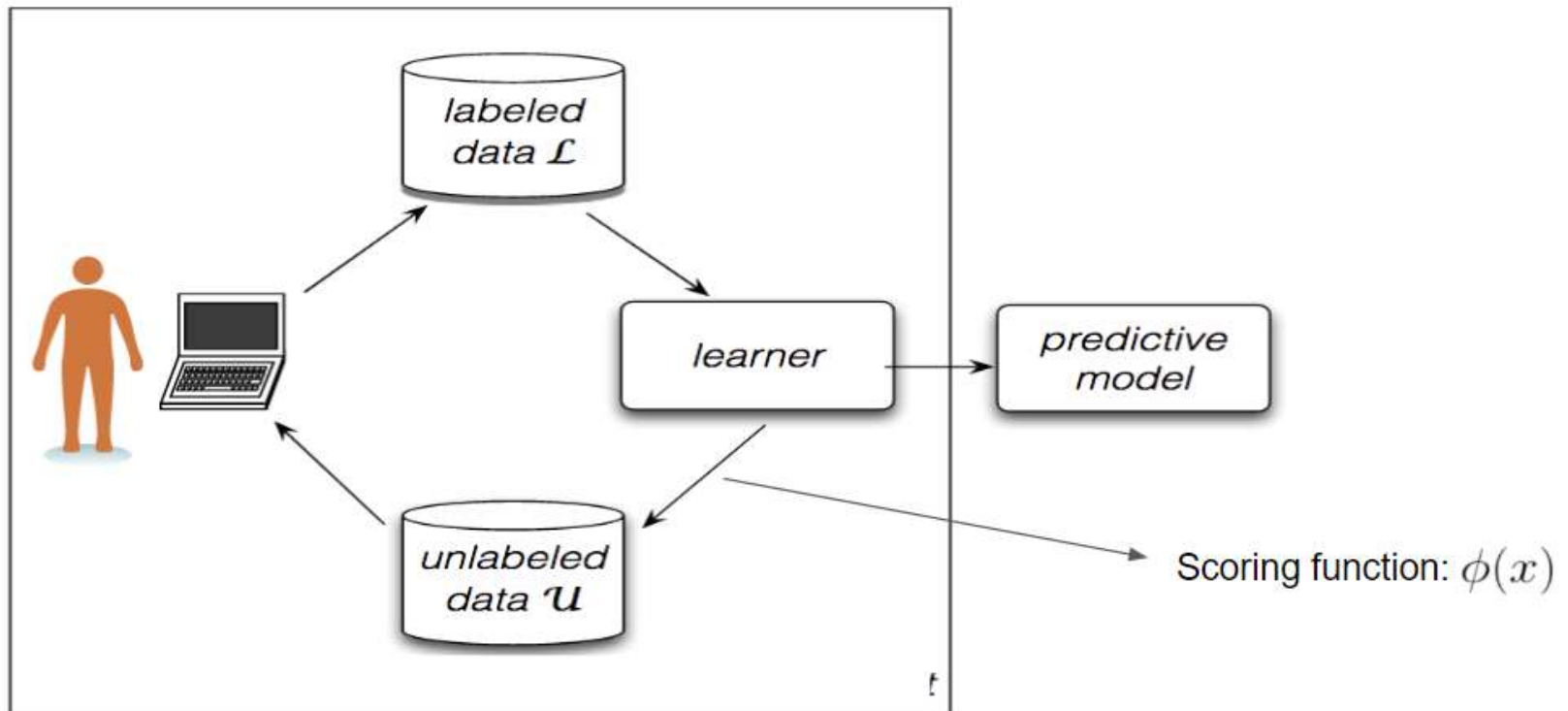




# Hybrid Man-Machine Relevance Judging

- Systematic review (medicine) and e-Discovery (law / civil procedure) have traditionally relied on trusted doctors/lawyers for judging
- Automatic relevance classification is more efficient but less accurate
- Recent active learning work has investigated hybrid man-machine judging combinations
  - e.g., TAR & TREC Legal Track, recent CLEF track

# What is Active Learning?



(Wallace et al., 2016)

---

# Information Retrieval and Web Search

## Introduction

# Relevance

---

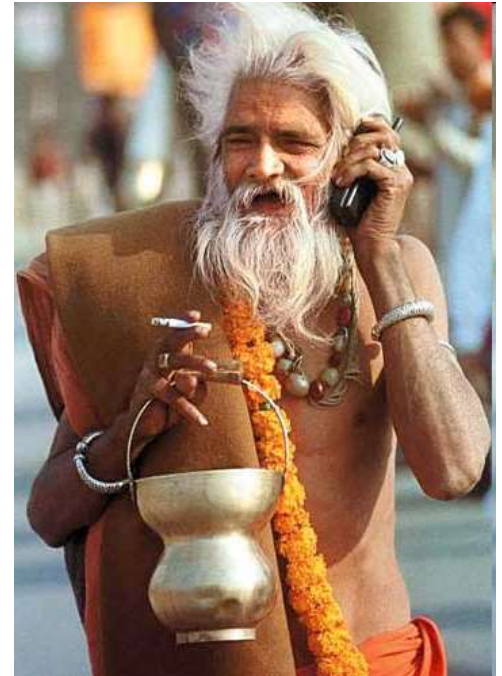
- Relevance is a subjective judgment and may include:
  - Being on the proper subject.
  - Being timely (recent information).
  - Being authoritative (from a trusted source).
  - Satisfying the goals of the user and his/her intended use of the information (*information need*).



# Relevance

- What is it?
  - **Simplistic definition**: A relevant document contains the information that a person was looking for when they submitted a query to the search engine
  - Many factors influence a person's decision about what is relevant: e.g., task, context, novelty, style
  - *Topical relevance* vs. *user relevance*

# Who and Where?



# Modeling Relevance

- *Ranking algorithms* used in search engines
- Ranking is typically statistical and based on its *observable* properties rather than *underlying* linguistic properties
  - i.e. counting simple text features such as words instead of inferring underlying linguistic syntax
  - However, both kinds of *features / evidence* can be incorporated into a statistical model

# Keyword Search

---

- Simplest notion of relevance is that the query string appears verbatim in the document.
- Slightly less strict notion is that the words in the query appear frequently in the document, in any order (*bag of words*).

# Problems with Keywords

---

- May not retrieve relevant documents that include synonymous terms.
  - “restaurant” vs. “café”
  - “PRC” vs. “China”
- May retrieve irrelevant documents that include ambiguous terms.
  - “bat” (baseball vs. mammal)
  - “Apple” (company vs. fruit)
  - “bit” (unit of data vs. act of eating)

# Users and Information Needs

- Search evaluation is user-centered
- Keyword queries are often poor descriptions of actual information needs
- Interaction and context are important for inferring user intent
- Query refinement techniques such as *query expansion*, *query suggestion*, *relevance feedback* improve ranking



# Query Disambiguation

- Given (typically terse like “apple”) query, infer possible underlying intents / needs / tasks
- With longer queries, detect key concepts and/or segment (e.g. “new york times square”)

The image shows a search engine results page for the query "apple". At the top, there is a search bar with "apple" entered and a "search" button. Below the search bar, it says "e.g. 'apple' or 'computer science'". The search results are displayed in a grid format. On the left side, there is a sidebar with a list of categories and their counts: AAPL Stock (24), Mac News (8), Apple Computer (7), Apple Online Store (5), Apple Developer Connection (5), and Apple Products (6). The main content area shows several search results, each with a title, a brief description, a URL, and a small image. The results include: AAPL Stock (24) with a link to "Digg - AAPL Stock Tanking" and a photo of a woman; Mac News (8) with a link to "Macworld" and images of an iPhone and a CD; Apple Computer (7) with a link to "Apple Investor Relations" and images of a laptop and a CD; Apple Online Store (5) with a link to "Apple Online Store" and images of a laptop and a CD; Apple Developer Connection (5) with a link to "Apple Developer Connection" and images of a laptop and a CD; and Apple Products (6) with a link to "Mac Laptops, iPods and Macintosh Computers" and images of a laptop and a CD.

**More Applications...**

# Location-based Search



**CYCLOPEDIA:**  
Augmented Wikipedia



1



See something you want to remember

When you notice an item to remember, tap "Remembers" in the Amazon App.

2



Snap photo & send

Your iPhone camera will open. Take a photo of the item and it will be sent to Amazon.



3



See reminders

Your photos & any similar products that Amazon finds are stored in the app and on Amazon.com.



The Amazon app includes **Amazon Remembers**

amazonmechanicalturk  
Artificial Artificial Intelligence



snaptell



# Content-based music search



# Retrieving Information, not Documents

**Correlator** from **YAHOO! RESEARCH**

haiti

 [Wikipedia](#)  [Names](#)  [Places](#)  [Events](#)  [Concepts](#)  [News](#)  [Answers](#)

### Events related to "haiti"

#### Timeline



- 1804: Haiti attempted to establish closer ties with the United States
- January 1, 1804: It is also known as Haïti's independence city

#### Events in the timeline

**January 1, 1804**

(From [W Gonaïves](#)) "It is also known as **Haïti's independence city** " because it was there that Gen. Jean-Jacques Dessalines declared **Haïti's** independence on **January 1, 1804** ."

(From [W Gonaïves](#)) "Gonaïves is also known as **Haïti's City of Independence** because it was there that Jean-Jacques Dessalines declared **Haïti** , the former Saint-Domingue , independent from France on **January 1, 1804** by reading the Act of Independence , drafted by Boisrond Tonnerre , on the Place d'Armes of the town ."

(From [W List of French Governors of Saint-Domingue](#)) "**January 1, 1804** , Independence of **Haïti** , Jean-Jacques Dessalines is Provisional Chief of the Haitian Government to September 22 , 1804 and then Emperor of **Haïti** until October 17 , 1806"

**1804**

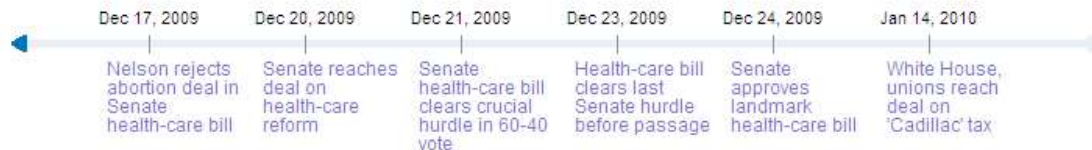
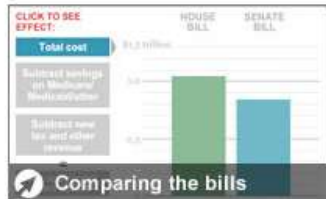



# News Tracking (*Living Stories*)

## Washington Tackles Health Care Reform

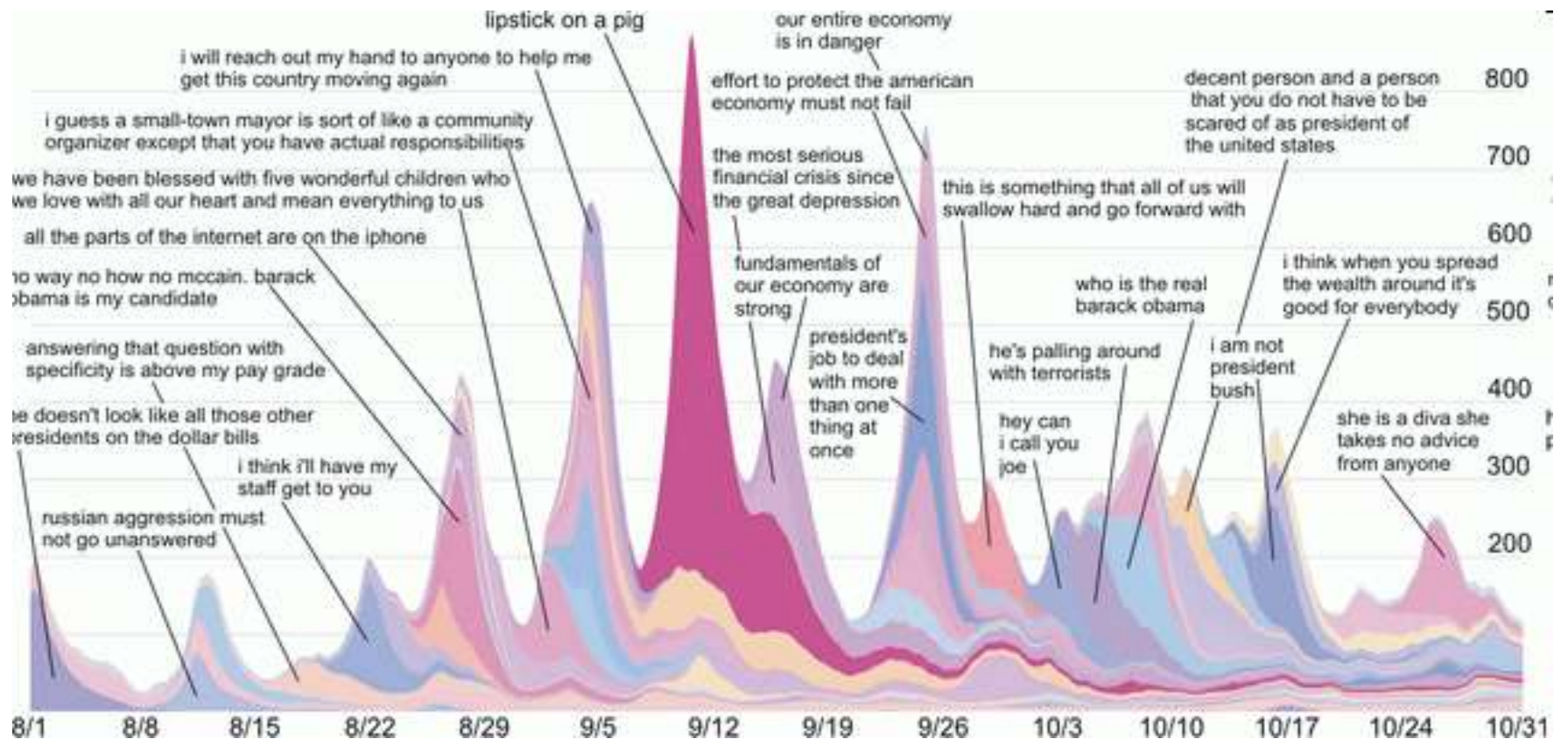
**Overview:** The House and Senate have approved sweeping legislation that would provide health care insurance for most Americans, at huge cost to the government. The House plan, approved Nov. 7 in an almost strictly party-line vote, would spend \$1.05 trillion to extend coverage to about 36 million Americans. The Senate bill, passed on Dec. 24, went through several iterations before attracting the support of a filibuster-proof coalition of 60 votes. It would cost \$871 billion and give coverage to 31 million people who lack it now.

[Read more...](#)



All Coverage	Obama stumps in Massachusetts with health-care reform at risk	11:17 AM	Timeline of important events
<a href="#">The cost</a> <a href="#">The politics</a> <a href="#">What's in the bills</a>	<p>By Karl Vick, Paul Kane</p> <p>The president tries to resurrect the candidacy of Democratic Senate candidate Martha Coakley in the surprisingly close race to succeed the late Edward M. Kennedy in Massachusetts. A Coakley loss to state Sen. Scott Brown (R) would deprive Democrats in the Senate of the super-majority they need to pass Obama's top domestic priority.</p> <p><b>Related</b></p> <p><a href="#">Obama pleads for pragmatism on health-care overhaul</a> - Feature</p> <p><a href="#">Read more...</a></p>	<p><a href="#">Jump to: People</a></p>	<p>White House, unions reach deal on 'Cadillac' tax Jan 14, 2010</p> <p>Senate approves landmark health-care bill Dec 24, 2009</p> <p>Health-care bill clears last Senate hurdle before passage Dec 23, 2009</p> <p>Senate health-care bill clears crucial hurdle in 60-40 vote Dec 21, 2009</p> <p>Senate reaches deal on health-care reform Dec 20, 2009</p>
<p>› All</p> <ul style="list-style-type: none"> <li>News</li> <li>Features</li> <li>Opinion</li> <li>People</li> <li>Resources</li> <li>Images</li> <li>Videos</li> <li>Graphics</li> </ul> <p>› Standard view</p> <p><a href="#">Most important only</a></p> <p>› Newest first</p> <p><a href="#">Oldest first</a></p>	<p><a href="#">White House, unions reach deal on 'Cadillac' tax</a></p> <p>By Lori Montgomery, Michael D. Shear</p> <p>A bargain on taxing high cost health insurance policies puts negotiators close to an overall deal on a health-care reform plan. Under the agreement, family plans that cost more than \$24,000 and individual policies that cost more than \$8,900 would be subject to a 40 percent surtax. The tax would be imposed on the insurance company, but economists believe it would be passed on to workers.</p> <p>Last year, the average family policy in America cost \$13,375, according to a survey by the Kaiser Family Foundation.</p> <p>The changes would cut revenue for the health reform package by \$60 million over 10 years, a sum likely to be made up by</p>	<p>Jan 14, 2010</p>	
	 <p>The White House cut a deal with organized labor on how to tax high-cost insurance policies.</p>		

# Memetracker



# “Hyper-local” Search

**EveryBlock** A news feed for your block

Track and discuss what's new in your neighborhood.

Choose a city...

- BETA Atlanta
- Boston
- Charlotte
- Chicago
- BETA Dallas
- BETA Detroit
- BETA Houston
- Los Angeles
- Miami
- New York
- Philadelphia
- San Francisco
- San Jose
- Seattle
- Washington, DC

Don't see your city?

**EveryBlock New York City**

Restaurant inspections

Restaurant inspections conducted recently by the New York City Department of Health and Mental Hygiene.

Welcome to EveryBlock's New York restaurant inspection section. Here, you can explore restaurant inspections in New York in various ways. Use "Search near an address" to find restaurant inspections near any specific block or neighborhood, and use the powerful custom filter to create your own reports according to various criteria.

Filter this data

Location

Search near an address:

Within: 8 blocks

Or choose a location:

- Boroughs...
- Community boards...
- Neighborhoods...
- Police precincts...
- ZIP codes...

Inspection date

Violation

- Non-food contact surface improperly constructed
- Not vermin-proof
- Misc
- Cold food held above 41°F (jamoked fish above 35°F) except during necessary preparation
- Improper plumbing

2,573 restaurant inspections

Dec. 13, 2008 – Jan. 12, 2009

By borough

Borough	Count	Percentage
Manhattan	1,130	44%
Brooklyn	477	19%
Queens	254	10%
The Bronx	190	7%
Staten Island	80	3%

By commu

Chicago

Accuracy: Good (167 feet)

Monday, April 13, 2009

Locations in the media

Time Out Chicago: Five things to do today: April 13

Chicago Foodies: Marzetti's Is Planks

Check out the free EveryBlock iPhone app.



## Find news near you

Enter an address and see nearby crime, news coverage, neighbor announcements and more.



## Browse news by topic

Get a citywide overview of news by category, then click on topics to find what you're looking for.



## Explore your city

Choose a neighborhood or ZIP code and see what's happening nearby or track trends over time.



# Recent IR History

---

- 2010's
  - Intelligent Personal Assistants
    - Siri
    - Cortana
    - Google Now
    - Alexa
  - Complex Question Answering
    - IBM Watson
  - Distributional Semantics
  - Deep Learning

# Deep (a.k.a. Neural) IR


# Growing Interest in “Deep” IR

- **Success of *Deep Learning* (DL) in other fields**
  - Speech recognition, computer vision, & NLP
- **Growing presence of DL in IR research**
  - e.g., SIGIR 2016 Keynote, Tutorial, & Workshop
- **Adoption by industry**
  - Bloomberg: [Google Turning Its Lucrative Web Search Over to AI Machines](#). October, 2015
  - WIRED: [AI is Transforming Google Search](#). The Rest of the Web is next. February, 2016.





# But Does IR *Need* Deep Learning?

- Chris Manning (Stanford)'s SIGIR Keynote:  "I'm certain that **deep learning will come to dominate SIGIR over the next couple of years**... just like speech, vision, and NLP before it."
- Despite great successes on short texts, **longer texts typical of ad-hoc search remain more problematic**, with only recent success (e.g., Guo et al., 2016)
- As Hang Li eloquently put it, "**Does IR (Really) Need Deep Learning?**" (SIGIR 2016 Neu-IR workshop)

# Neural Information Retrieval: A Literature Review

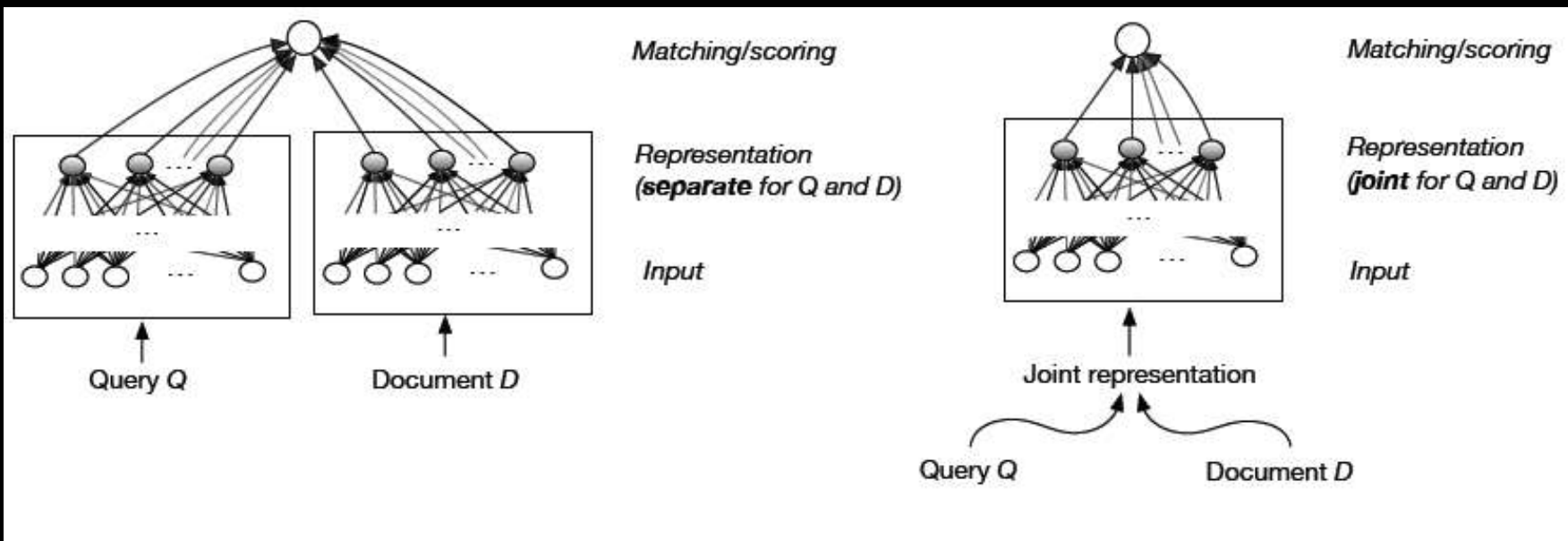
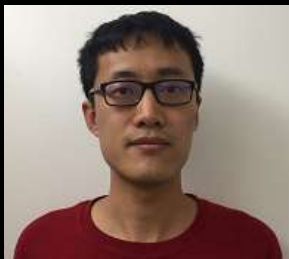


Figure 4: Two basic neural architectures for scoring the relevance of queries to documents.



Ye Zhang et al.

<https://arxiv.org/abs/1611.06792>

Posted 18 November, 2016



# Word Embeddings



@mattlease

# Traditional “one-hot” word encoding

Leads to famous *term mismatch* problem in IR

## The standard word representation

The vast majority of rule-based **and** statistical NLP work regards words as atomic symbols: *hotel, conference, walk*

In vector space terms, this is a vector with one 1 and a lot of zeroes

$[0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 1 \ 0 \ 0 \ 0 \ 0]$

Dimensionality: 20K (speech) – 50K (PTB) – 500K (big vocab) – 13M (Google 1T)

We call this a “one-hot” representation. Its problem:

*motel*  $[0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 1 \ 0 \ 0 \ 0 \ 0]$  AND  
*hotel*  $[0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0 \ 1 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0]$  = 0

# Distributional Representations

Define words by their co-occurrence signatures

You can get a lot of value by representing a word by means of its neighbors

“You shall know a word by the company it keeps”

(J. R. Firth 1957: 11)

One of the most successful ideas of modern statistical NLP

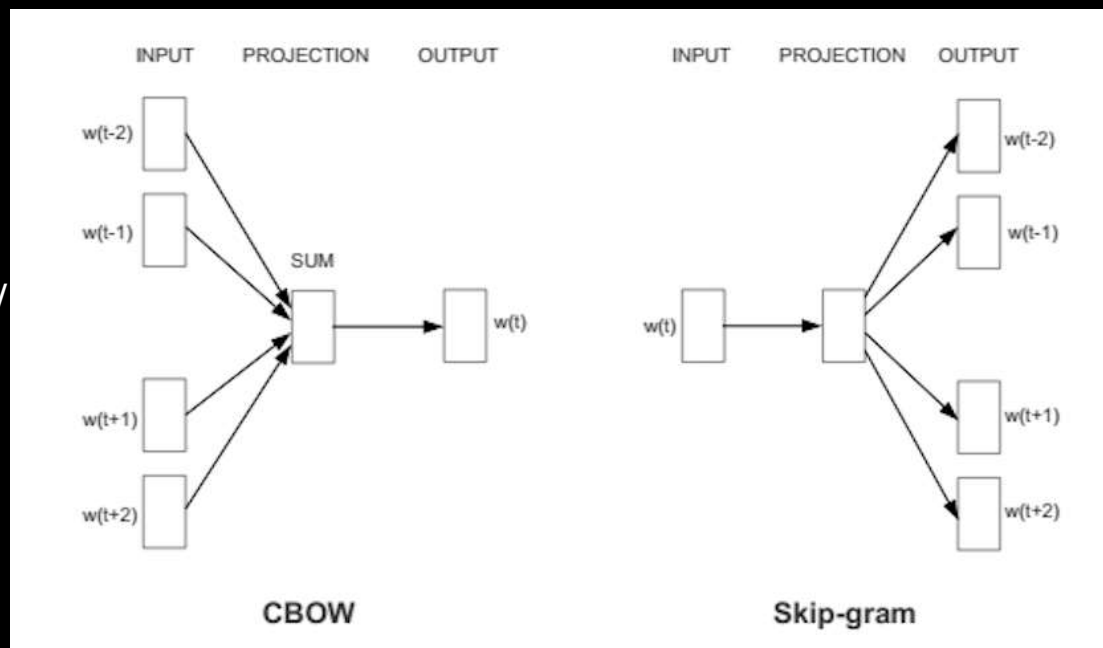
government debt problems turning into banking crises as has happened in  
saying that Europe needs unified banking regulation to replace the hodgepodge

↖ These words will represent *banking* ↗

You can vary whether you use local or large context to get a more syntactic or semantic clustering

# “Early” Neural *Word Embeddings*

- **word2vec (Mikolov et al., 2013)** – sliding window
  - CBOW: predict center word given window context
  - Skip-gram: predict context given center word



[deeplearning4j.org/  
word2vec](http://deeplearning4j.org/word2vec)

- See also: **GloVe (Pennington et al., 2014)**



# Extending IR Models with Word Embeddings



@mattlease

# Recent IR Work with Word Embeddings

Task	Studies
Ad-hoc Retrieval	ALMasri et al. (2016), Amer et al. (2016), BWESG (Vulic and Moens (2015)), Clinchant and Perronnin (2013), Diaz et al. (2016), GLM (Ganguly et al. (2015)), Mitra et al. (2016), Nalisnick et al. (2016), NLTM (Zuccon et al. (2015)), Rekabsaz et al. (2016), Roy et al. (2016), Zamani and Croft (2016a), Zamani and Croft (2016b), Zheng and Callan (2015)
Bug Localization	Ye et al. (2016)
Contextual Suggestion	Manotumrukxa et al. (2016)
Cross-lingual IR	BWESG (Vulic and Moens (2015))
Detecting Text Reuse	Zhang et al. (2014)
Domain-specific Semantic Similarity	De Vine et al. (2014)
Community Question Answering	Zhou et al. (2015)
Short Text Similarity	Kenter and de Rijke (2015)
Outlier Detection	ParagraphVector (Le and Mikolov (2014))
Sponsored Search	Grbovic et al. (2015b), (Grbovic et al., 2015a)

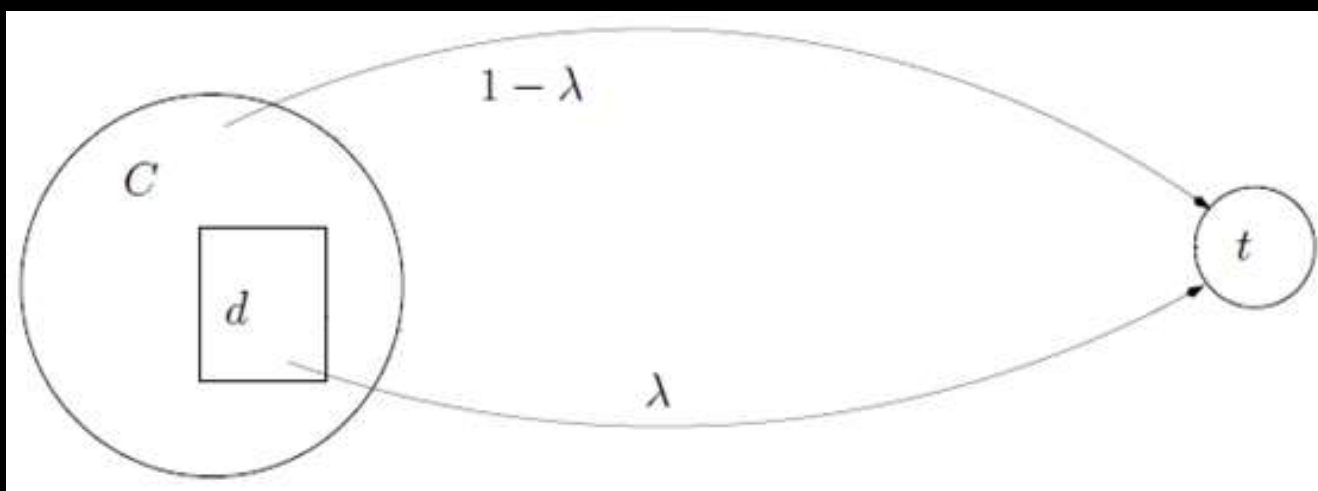
# Ponte & Croft (2001): LM for IR

$$P(D|Q) = [ P(Q|D) P(D) ] / P(Q)$$

$$\propto P(Q|D) P(D) \quad \text{for fixed query}$$

$$\propto P(Q|D) \quad \text{assume uniform } P(D)$$

$$P(Q|D) = \prod_q \alpha * P(q|D) + (1 - \alpha)P(q|C)$$

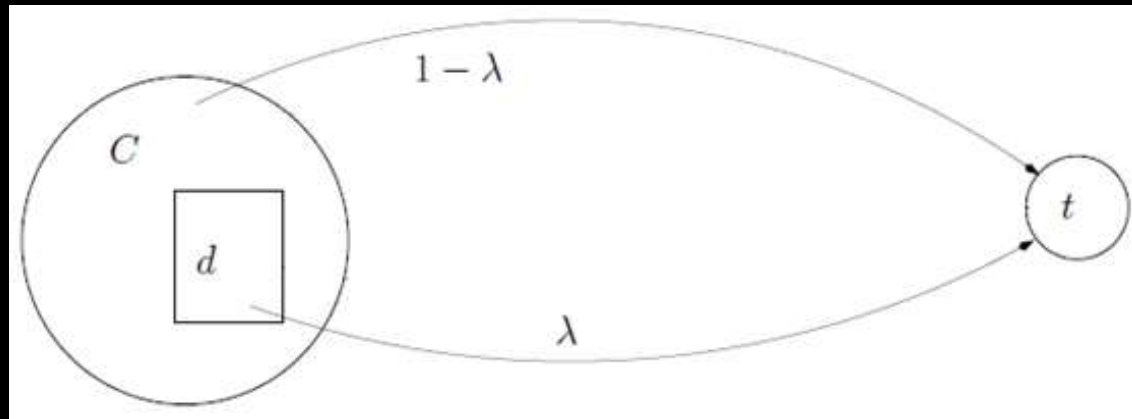


# Berger & Lafferty (1999)

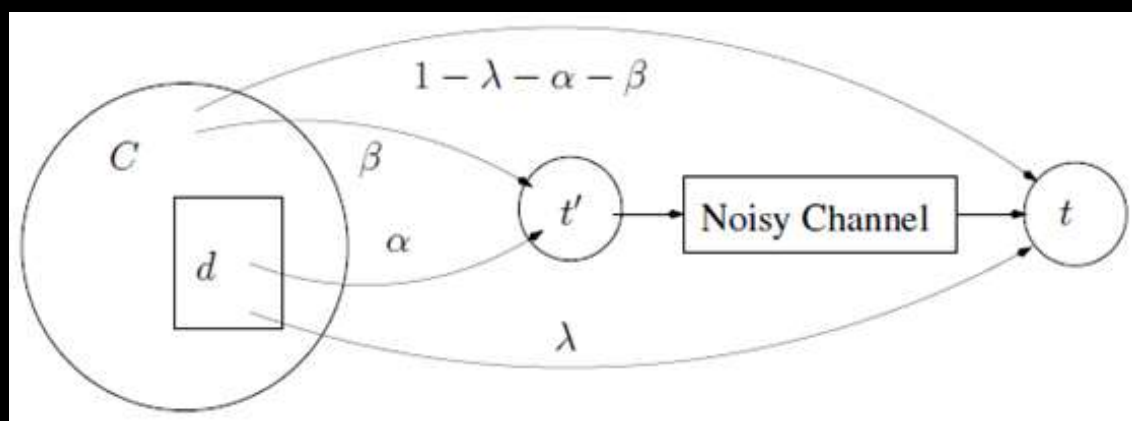
- IR as Statistical Translation
  - Document  $d$  contains word  $w$
  - $w$  is translated to observed query word  $q$

$$\begin{aligned} p_\alpha(q | \mathbf{d}) &= \alpha p(q | \mathcal{D}) + (1 - \alpha) p(q | \mathbf{d}) \\ &= \alpha p(q | \mathcal{D}) + (1 - \alpha) \sum_{w \in \mathbf{d}} l(w | \mathbf{d}) t(q | w) \end{aligned}$$

# GLM: Ganguly et al., SIGIR 2015



# GLM: Ganguly et al., SIGIR 2015



$$P(t|t', d) = \frac{\text{sim}(t, t')}{\sum_{t'' \in d} \text{sim}(t, t'')} = \frac{\text{sim}(t, t')}{\Sigma(d)}$$

$\text{sim}(t, t')$  is the cosine similarity between the vector representations of  $t$  and  $t'$  and  $\Sigma(d)$  is the sum of the similarity values between all term pairs



# NTLM: Zuccon et al., ACDS 2015

Berger  
and Lafferty have proposed an alternative estimation of  $p_s(w|d)$   
inspired by models in statistical machine translation [5].

$$p_t(w|d) = \sum_{u \in d} p_t(w|u)p(u|d)$$

## Estimating Translation Probabilities with Neural Language Models

The use of neural language models based on continuous bag-of-words or skipgram gives rise to two different word embeddings. Word embeddings can be used to estimate translation probabilities in translation language models; specifically, cosine similarity can be used as a proxy for  $p(u|w)$ :

$$p_{cos}(u|w) = \frac{\cos(u, w)}{\sum_{u' \in V} \cos(u', w)}$$

# DeepTR: Zheng & Callan, SIGIR 2015

- Supervised learning of effective term weights
  - Like *RegressionRank* (Lease et al., ECIR 2009), (Lease, SIGIR 2009) but without feature engineering
- Represent each query term in context by avg. query embedding - term embedding

DeepTR-BOW:

```
#weight(  $\hat{P}(\text{apple}|R)$  apple  
           $\hat{P}(\text{pie}|R)$  pie  
           $\hat{P}(\text{recipe}|R)$  recipe )
```

DeepTR-SD:

```
#weight(  
0.8 #weight(  $\hat{P}(\text{apple}|R)$  apple  
              $\hat{P}(\text{pie}|R)$  pie  
              $\hat{P}(\text{recipe}|R)$  recipe )  
0.1 #combine(#1(apple pie)  
             #1(pie recipe) )  
0.1 #combine(  
             #uw8(apple pie)  
             #uw8(pie recipe) ) )
```

# Diaz, Mitra, & Craswell, ACL 2016

- Learn topical word embeddings *at query-time*
  - New flavor of classic IR *global vs. local* tradeoff
  - Compare use of collection vs. external corpora
- No comparison to pseudo-relevance feedback

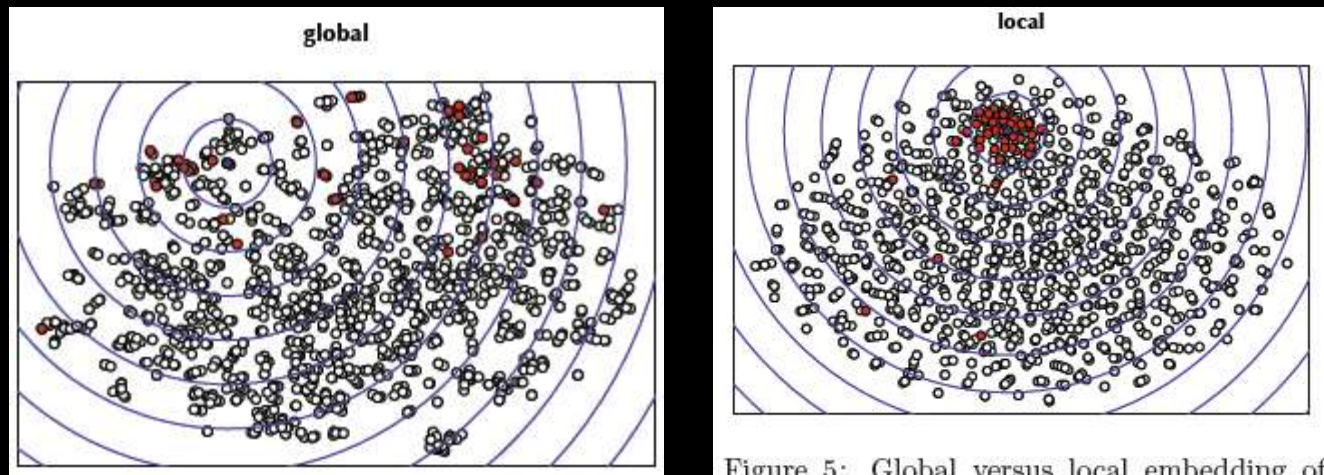


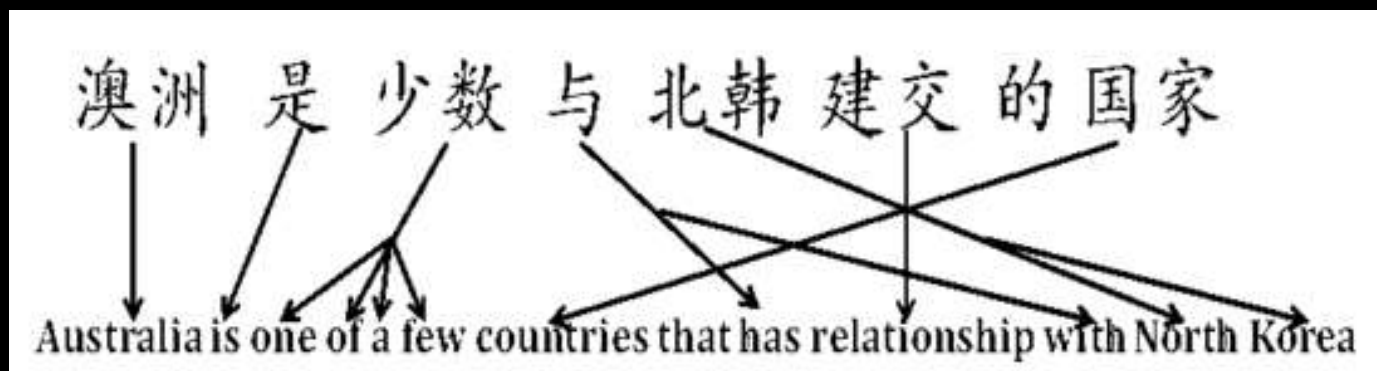
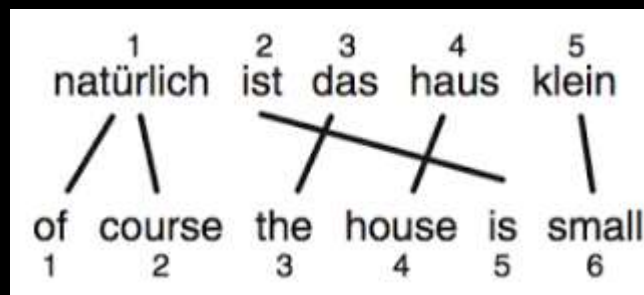
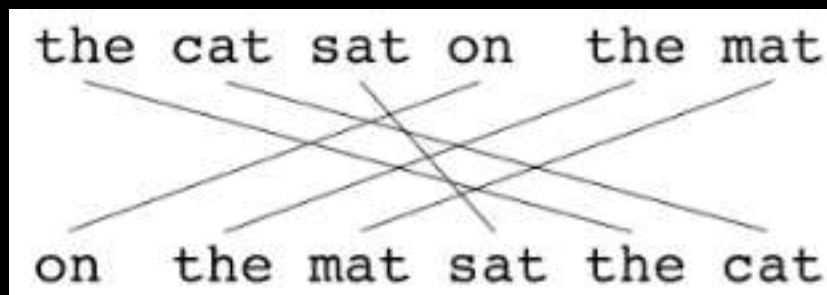
Figure 5: Global versus local embedding of highly relevant terms. Each point represents a candidate expansion term.

# Cross-Lingual IR with Bilingual Word Embeddings



# Bilingual Word Embeddings for Phrase-Based Machine Translation

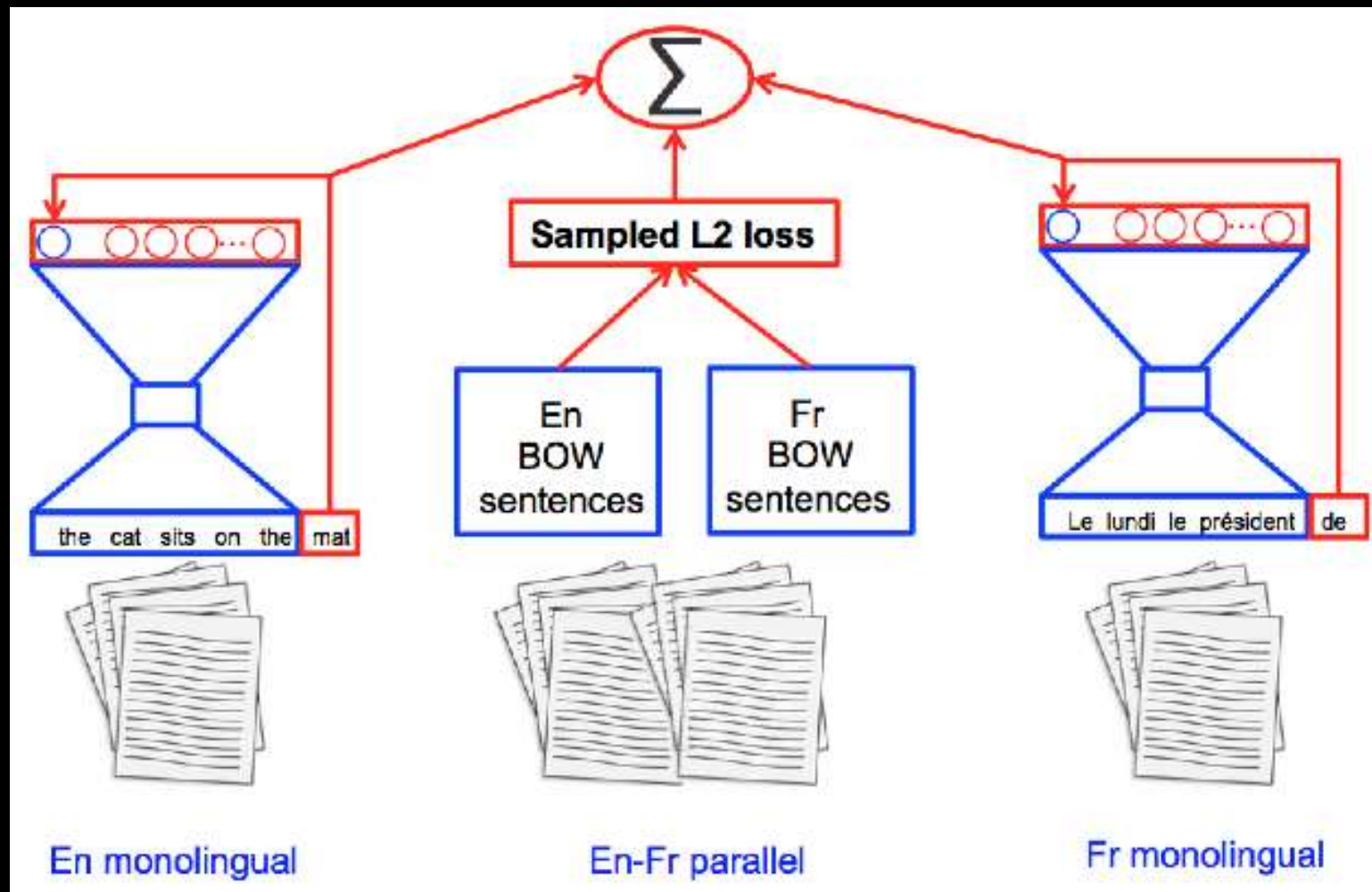
Will Y. Zou<sup>†</sup>, Richard Socher, Daniel Cer, Christopher D. Manning  
Department of Electrical Engineering<sup>†</sup> and Computer Science Department  
Stanford University, Stanford, CA 94305, USA





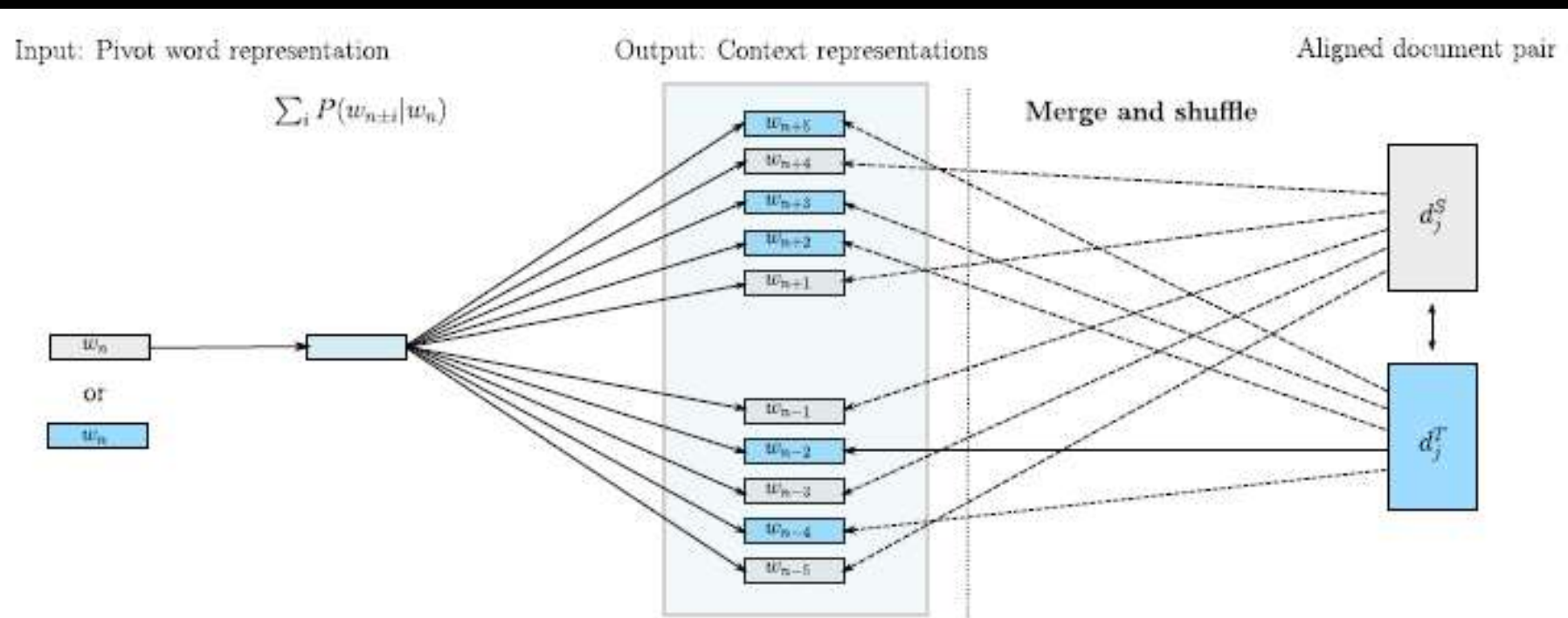
# BilBOWA: Fast Bilingual Distributed Representations without Word Alignments

[Stephan Gouws](#), [Yoshua Bengio](#), [Gregory S. Corrado](#) • ICML • 2015





# BWESG: Vulic & Moens, SIGIR 2015



**Merge & Shuffle:** Training a SGNS (or any other monolingual model!) on shuffled "pseudo-bilingual" documents

# Ye et al., ICSE 2016: Finding Bugs

- Given textual bug report (query), find software files needing to be fixed (documents)
  - Saha, Lease, Khursid, Perry (ASE, 2013)
- Augment the Skip-gram model to predict all code tokens from each text word, and all text words from each code

```
void connect(IStreamsProxy streamsProxy)
```

Connects this console to the given streams proxy. This associates the standard in, out, and error streams with the console. Keyboard input will be written to the given proxy.

Figure 9: Example of semantically related text and code, from API documents.

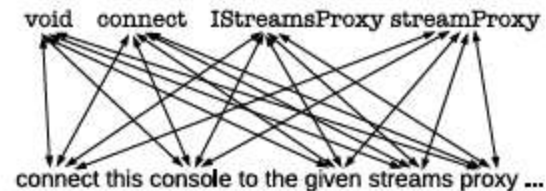


Figure 10: Positive pairs generated from semantically related text and code.

# Going Deeper with Characters

arXiv.org > cs > arXiv:1606.01781

Computer Science > Computation and Language

## Very Deep Convolutional Networks for Natural Language Processing

Alexis Conneau, Holger Schwenk, Loïc Barrault, Yann Lecun

*(Submitted on 6 Jun 2016)*

“The dominant approach for many NLP tasks are recurrent neural networks, in particular LSTMs, and convolutional neural networks. However, **these architectures are rather shallow in comparison to the deep convolutional networks which are very successful in computer vision.**

We present a new architecture for text processing which **operates directly on the character level** and uses only small convolutions and pooling operations. We are able to show that the performance of this model increases with the depth: **using up to 29 convolutional layers**, we report significant improvements over the state-of-the-art on several public text classification tasks. To the best of our knowledge, **this is the first time that very deep convolutional nets have been applied to NLP.**”

# Resources

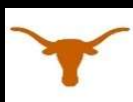
<http://deeplearning.net>



@mattlease

# Neural IR Source Code Released

System	Citation	URL
word2vec	<a href="#">Mikolov and Dean (2013)</a>	<a href="https://code.google.com/archive/p/word2vec/">https://code.google.com/archive/p/word2vec/</a>
GloVe	<a href="#">Pennington et al. (2014)</a>	<a href="http://nlp.stanford.edu/projects/glove/">http://nlp.stanford.edu/projects/glove/</a>
CDNN	<a href="#">Severyn and Moschitti (2015)</a>	<a href="https://github.com/aseveryn/deep-qa">https://github.com/aseveryn/deep-qa</a>
DeepMerge	<a href="#">Lee et al. (2015)</a>	<a href="https://ciir.cs.umass.edu/downloads/DeepMerge/">https://ciir.cs.umass.edu/downloads/DeepMerge/</a>
DeepTR	<a href="#">Zheng and Callan (2015)</a>	<a href="http://www.cs.cmu.edu/~gzheng/code/TermRecallKit-v2.tar.bz2">http://www.cs.cmu.edu/~gzheng/code/TermRecallKit-v2.tar.bz2</a>
Mixed Deep	<a href="#">Gupta et al. (2014)</a>	<a href="http://www.dsic.upv.es/~pgupta/mixed-script-ir">http://www.dsic.upv.es/~pgupta/mixed-script-ir</a>
NTLM	<a href="#">Zuccon et al. (2015)</a>	<a href="https://github.com/ielab/adcs2015-NTLM">https://github.com/ielab/adcs2015-NTLM</a>



# Thank You!



**Slides:**

[slideshare.net/mattlease](https://www.slideshare.net/mattlease)

**Lab:** [ir.ischool.utexas.edu](http://ir.ischool.utexas.edu)

